5 The case for physicalism

This chapter will be concerned to argue for, to elaborate, and to defend physicalism. Physicalists maintain that all of the states and processes involved in the human mind are, at bottom, physical states and processes. Since physicalism is the denial of weak dualism ('mental states are non-physical states'), when dualism is referred to in this chapter it will be the weak version which is in question. And if weak dualism is rejected, then so too, of course, must strong dualism be rejected. If mental states themselves are physical, then the *subject* of those states surely couldn't be *non*-physical.

1 Arguments for mind–brain identity

What the thesis of mind–brain identity affirms is that descriptions of our mental states, on the one hand, and some descriptions of our brain states, on the other, are in fact descriptions of the very same things. It holds that just as a particular cloud is, as a matter of fact, a great many water droplets suspended close together in the atmosphere; and just as a flash of lightning is, as a matter of fact, a certain sort of discharge of electrical energy; so a pain or a thought is (is identical with) some state of the brain or central nervous system.

The identity-thesis is a version of physicalism: it holds that all mental states and events are in fact physical states and events. But it is not, of course, a thesis about meaning: it does not claim that words such as 'pain' and 'after-image' may be *analyzed* or *defined* in terms of descriptions of brain-processes. (That would be absurd.) Rather, it is an empirical thesis about the things in the world to which our words refer: it holds that the ways of thinking represented by our terms for conscious states, and the ways of thinking represented by some of our terms for brain-states, are in fact different ways of thinking of the very same (physical) states and events. So 'pain' doesn't *mean* 'such-and-such a stimulation of the neural fibers' (just as 'lightning' doesn't *mean* 'such-and-such a discharge of electricity'); yet, for all that, the two terms in fact refer to the very same thing.

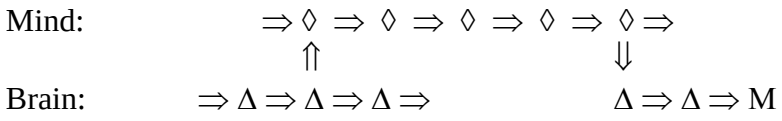
In this section a number of arguments in support of mind–brain identity will be set out and discussed. All of these arguments are broadly empirical ones, drawing on our beliefs about the causal order of the world, and our place within it.

1.1 The closure of physics and the unity of nature

Almost everyone believes that mind and matter interact causally with one another. For example, stimulation of our sense-organs causes conscious experiences, and decisions cause bodily movements. An interactive-dualist will then have to picture the situation somewhat as represented in figure 5.1. (The diamonds here represent mental events, and the triangles represent physical ones; the arrows represent causality, and 'M' represents a bodily movement of some

sort. Notice that on this picture there is a brain event amongst the causes of movement M *which has no physical cause*. This is a point we will return to in section 1.2 below. Notice, too, that the diagram is, of course, hugely over-simplified. Normally many different mental states will contribute to the causation of a given mental state, and many different complex patterns of brain states will contribute to the causation of any given brain state or bodily movement.)

Figure 5.1: Interactive dualism



One of the main objections to dualism has always been the difficulty of making sense of causal connections between mind and brain, as we saw in chapter 2:2. Now, there isn't any problem of *principle* in understanding causal connections between physical and non-physical realms (as we argued in chapter 2:2). For there is nothing in the concept of causation, as such, which requires all that causes be mediated by physical mechanisms. The real problem is to understand *how* such causation can occur, given what we already know or believe about the physical world, and about causation in the brain.

Consider, first, the physical world in general. Most scientists now believe that *physics is closed*, in the sense of permitting no interference from, or causation by, events at higher levels of description (e.g. chemical or biological). On this view, all atomic and sub-atomic events happen in accordance with physical laws (albeit probabilistic ones), and all events at higher, more abstract, levels of description must be realized in, or constituted by, those physical processes, in such a way as to allow no independent point of causal leverage. So while there may be chemical and biological laws, the events which figure in these laws must always, at the same time, fall under the laws of physics. (I shall say some more about this in section 2.2 below.)

On this conception, there is simply *no room* for a distinct and independent psychological level of nature, whose events are not physically constituted, but which can have an impact upon the physical behavior of the body. For in order for such a thing to be possible, it would have to be the case that non-physical mental events could have an impact on causal sequences at the physical level. But this would conflict with the causal closure of physics – it would mean that some physical events would be caused, not by other physical events or processes, but rather by non-physical mental events.

What reason do we have for believing in the causal closure of physics? This is not something which can be proved (least of all by thought alone, of course). But for some centuries it has been a successful methodological assumption of scientific enquiry. Scientists work under the assumption that processes in physics brook no interference from higher levels of causation. And whenever they come across physical phenomena which cannot presently be explained in

physical terms, instead of postulating causation by *élan vital* (a supposed independent biological life-force), or causation by *ectoplasm* (a supposed independent psychic force), or whatever, they look deeper into the physical mechanisms. In many such cases this deeper look has proved successful; and in all such cases physicalistic scientific enquiries continue to make progress. This gives us good reason to think that the scientific methodology is correct, and that physics is indeed closed.

Closely related to the principle of the causal closure of physics is the principle of the *unity of nature*. On this conception, nature is *layered* into a unitary system of laws and patterns of causal organization, with the processes in any given layer being realized in the one below it. The bottom layer is fundamental physics, which *realizes* (or *constitutes*) all the rest. Chemical laws and processes are realized in those of atomic physics, bio-chemical processes are constituted by those of molecular chemistry, biological and neurological processes are realized in those of bio-chemistry, and so on. (Again, more on this in section 2.2 below.) In accordance with this layered picture of nature, we should expect the principles and processes of human psychology – or the 'laws' of operation of the human mind – to be realized in those of neurology. That is to say: we should expect mental events to be constituted by physical events in the brain.

The basic reason for believing in the unity of nature (like our reason for believing in the closure of physics) is that it is a highly successful working methodological assumption of much scientific enquiry. Although scientists are concerned to discover the laws and principles which operate at any given level of organization in nature – biological, say – it is also an important goal of science to try to understand how those same laws might be *constituted* or *realized* by patternings of events at lower levels. They seek to understand how the right sequences of events at the lower level – that of bio-chemistry, say – would give rise to the patterns observed at the higher one. (When successful, the result is a *reductive explanation* of the higher-level phenomenon. The difference between reduction and reductive explanation is discussed in section 2.2 below.) This methodological assumption, too, has proved immensely successful, giving us reason to believe that mental processes will somehow be constituted by processes in the brain.

These arguments may be summarized as follows:

(1) It is a successful methodological assumption of science that non-physical events cannot cause physical ones. (*The closure of physics*.)

(2) It is a successful methodological assumption of science that higher-level events and processes in nature must be realized in lower-level (ultimately physical) ones. (*The unity of nature*.)

(C) So we have reason to think that mental events must be realized in physical ones, probably in physical events in the brain.

The argument is broadly inductive in form, since it projects forward from the success of assumptions made by scientists in the past to a new case. But it seems none the worse for that. For in the light of our endorsement of Empiricism and rejection of Rationalism in chapter 4, we

should in any case be wary of attempts to *prove* the truth of physicalism. On the contrary, good inductive arguments are just what an Empiricist might be expected to look for.

1.2 The argument from causation in the brain

Now consider, more particularly, what we believe about the nature of the causal processes which take place in the human brain. There is much still to learn about the brain – about the functions and interactions of its parts, for example. But much is already known. It is known that the brain consists of nerve cells, of various known types. And much is known about how such cells function, and the physical causes which lead to their activity. Certainly there appear to be no 'inverse causal black-holes' in the brain, such as would seem to be required by the interactionist picture. (That is, there are no places in the brain where brain activity begins to occur *for no physical reason*.) Indeed, I claim that enough is already known about the brain to justify the following principle: *each event in the brain has a sufficient physical cause*. In our picture, then, the chain of events in the brain leading to any given bodily movement ought to be *unbroken*, as in figure 5.2. How can this be made consistent with interactive dualism?

Figure 5.2: Causation in the brain

Brain: $\Rightarrow \Delta \Rightarrow \Delta \Rightarrow \Delta \Rightarrow \Delta \Rightarrow \Delta \Rightarrow \Delta \Rightarrow$ M

As we noted in chapter 2:2, we believe very firmly that some mental states and events are causally necessary for the occurrence of some physical ones. For example, I believe that if I had not been conscious of a pain in my foot (mental event), I should not have gone to the doctor (physical event). My awareness of the pain was, I believe, a causally necessary condition of my later visit to the doctor. But as we noted above, it seems most unlikely that we shall ever need to advert to anything other than physical–physical causality when we investigate the detailed causal nexus behind any given bodily movement. On the contrary, it seems likely that there will always be physical events providing us with a sufficient causal explanation of the brain events giving rise to any particular bodily movement. For example, as we trace the causes of my legs moving me in the direction of the doctor's surgery, through events in the muscles of my legs and feet, through events in the nerves of my spinal column, into events in the cells of my brain, it seems most unlikely that the chain of physical causation will eventually run out. Indeed, according to the principle set out above, we already know enough about causation in the brain to know that it won't. So we shall never be forced to appeal to any non-physical event in order to provide a satisfactory causal explanation of the movements leading to my visit to the surgery.

Now the only way in which we can hold onto both beliefs – the belief that some mental events are causally necessary for the occurrence of some physical ones, and the belief that it is unnecessary to appeal to anything other than physical events in providing causal explanations of brain events – is by believing that some mental events *are* physical ones. Then somewhere in the chain of physical causes of my visit to the doctor there will be a brain event which is (is identical

with) my awareness of a sensation of pain.

This argument for the general truth of the mind–brain identity-thesis may be summarized as follows.

(1) Some mental states and events are causally necessary for the occurrence of some physical ones.

(2) In a completed neuro-physiological science there will be no need to advert to anything other than physical–physical causality in the brain.

(C) So some mental states and events are (are identical with) physical (brain) states and events.

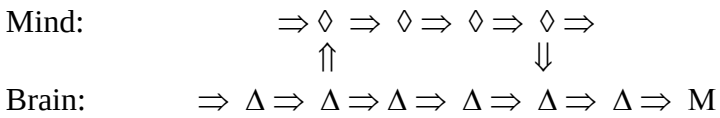
The argument is valid. And although its conclusion only claims that *some* mental states are physical, it can easily be developed in such a way as to entail the stronger conclusion that all are. For almost every kind of mental state can sometimes be causally necessary for a physical one, we think. Sometimes a particular bodily movement would not have taken place if I had not made a particular decision; or if I had not entertained a particular thought; or if I had not been aware of a particular sensation; or if I had not had a particular after-image; and so on. Then since it seems extremely unlikely that some mental states are physical while some are not, it follows that all are.

1.3 Causal over-determination

An interactive-dualist may try to get around the difficulty by appealing to the notion of 'causal over-determination'. Very roughly, this is the idea that an event may have more causes than are necessary. For example, imagine someone being shot by a firing squad, each member of which has a loaded gun (contrary to normal practice). Suppose that every soldier's aim is true, that all fire at the same time, and that every bullet strikes the heart. Then it is true of every soldier, that even if the others had not fired, his action would still have caused the prisoner's death. (Each shot individually is causally sufficient for death.) But it is also true of every soldier, that even if he himself had *not* fired, the prisoner's death would still have been caused by the others. (No shot individually is causally necessary for death.) Similarly then: the dualist may propose that brain-events are caused *both* by prior brain-events (so the chain of physical causes is unbroken) *and* by prior mental events; where either type of event on its own is sufficient to produce the effect, but neither type of event on its own is necessary.

So we have: each shot is *causally sufficient* for death, in that death will follow from it, in the circumstances, even if the other shots aren't fired; but no shot is (individually) *causally necessary*, since even if it isn't fired, death will still be caused by the other shots. Perhaps, similarly, the events in the brain which cause bodily movements are caused *both* by earlier brain events *and* by certain mental events, such as a decision. The dualist's resulting conception of the relationship between mind and brain can then be schematized as in figure 5.3 (where the arrows now represent causal sufficiency).

Figure 5.3: Causal over-determination



By deploying the thesis of causal over-determination, a dualist can hold onto one aspect of our common-sense beliefs in face of the likely discovery of unbroken causal chains of brain-events. Namely: the belief that our decisions are sometimes, in the circumstances, *sufficient* to bring about a bodily movement. Yet one aspect of common-sense would still have to be given up. Namely: the belief that a decision is sometimes causally *necessary* for a bodily movement to occur.

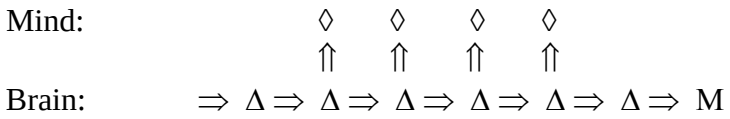
People who take the over-determination view can believe the following: given that a subject is sitting at a keyboard and decides to start typing, then this is, in the circumstances, sufficient for typing to begin. But they can no longer claim that had the subject *not* decided in that way, then the bodily movement wouldn't have taken place. On the contrary, it would still have occurred, brought about by its other cause: a particular brain-event. But are we really prepared to give up this belief? Don't I believe almost as firmly as I believe anything, that if I had not decided to write this book (mental event) I should not now be typing at this keyboard (physical event)?

How, then, can a dualist explain the fact that decisions are causally necessary *and* sufficient for bodily movements, consistent with our beliefs about the brain?

1.4 Epiphenomenalism

One suggestion is that we should give up believing that our decisions make any real causal difference. Rather, those decisions are mere *epiphenomena*, produced as a by-product by the brain events which are the true causes of our actions. Our picture of the relation between mind and brain will then be as represented in figure 5.4.

Figure 5.4: Epiphenomenalism



An attractive feature of this account is that it can explain how we come to be under the *illusion* of agency, falsely believing that our decisions are *causally* necessary and sufficient for some of our movements. For a given mental event will, on this account, be *non-causally* necessary and sufficient for a given movement, since each of them has a common cause. In fact the epiphenomenalist-dualist can claim that each mental event will be correlated with a particular brain-event as a matter of causal necessity. In which case it will be causally impossible for the

mental event to occur without the corresponding brain-event occurring.

So, given that the mental event occurs, then so too must the action occur which is caused by the underlying brain-event. (That is to say, the mental event is non-causally sufficient for the occurrence of the action.) And if the mental event *hadn't* occurred, then that would mean that the brain event hadn't occurred either, and so nor would the movement happen. (That is to say, the mental event is non-causally necessary for the occurrence of the action.) On this account, then, it will be *true* that if I hadn't decided to write this book, I shouldn't now be typing. For the only way in which I could have failed to take that decision, would have been if the corresponding brain-event had failed to occur; and if that had failed to occur, then the bodily movement wouldn't have been caused.

Compare the froth on the wave which breaks a sand-castle on the beach (see figure 5.5). Supposing that it is a law of nature that breaking waves produce froth on their leading edge, then we can say this: if the froth had not been there, the sand-castle wouldn't have been broken; and given that the froth is there, the sand-castle must be broken shortly thereafter. But it isn't really the froth which causes the sand-castle to break; rather, it is the wave which causes *both* the frothing *and* the breaking.

Insert *Figure 5.5 – The wave and the sandcastle* about here

Similarly we can say: if I hadn't decided to type, then I wouldn't now be typing. (This is because, if the event of my deciding hadn't occurred, then that would have been because the brain-event which caused my movement hadn't occurred.) And we can say: given that I decide to type, in the circumstances, then typing occurs. (This is because, if the decision to type occurs, then that will have been caused by the brain event which causes the relevant movements.) But my decision won't be the true cause of my typing, any more than the froth on the wave is the true cause of the sand-castle breaking.

Although a theory of this sort can save our belief that certain of our bodily movements wouldn't have occurred if certain decisions hadn't been taken, it does so at the cost of explanatory redundancy. For the decision is no longer part of the true causal explanation of why the bodily movement took place. To say that our decisions are causally correlated with the events which cause our bodily movements, isn't the same as saying (what we intuitively believe) that our decisions themselves constitute the true causal explanations of our actions.

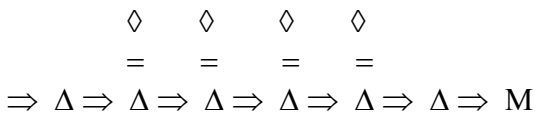
It seems that epiphenomenalism must conflict with our common-sense belief in the reality of agency – it conflicts with our belief that our decisions can make a causal difference to what we do. But another reason why epiphenomenalism is unacceptable is that, if it were true, it would remain a complete mystery why our decisions should march so neatly in step with our actions. Why is it that the brain event which causes me to sit down to type *also* causes me to think, 'Now I will begin typing'? For, by hypothesis, it wouldn't have made the slightest bit of difference if that brain event had caused me to think instead, 'Now I'll go swimming'.

How would the underlying causal properties of the brain ever have evolved, for example? What would be the advantage if the brain event which causes my arm to go up also causes me to decide, 'Now I shall raise my arm'? For, by hypothesis, the latter has no causal effects in its own right. What difference would it have made if the brain event had caused me to decide, 'Now I shall open my mouth' or, 'Now I shall sit down' instead? It would seem to be a quite remarkable cosmic coincidence that the evolutionary processes which caused our brains to have their causal powers in respect of bodily movement, *also* led them to cause content-relevant mental events. Then since it is good explanatory practice to minimize miracles, we have good reason to reject epiphenomenalism, and to endorse the thesis of mind–brain identity instead.

1.5 Mind–brain identity

In fact our belief in the reality of agency (that is, our belief that our decisions are often part of the true causal explanation of our actions) is very deeply held. If we are to give it up, then there had better be some powerful arguments for dualism to force us to do so. But in fact, as has emerged from chapters 2 and 4, there are none. The only remaining picture of the relation between mind and brain, then, is one of *identity*, as represented in figure 5.6. On this account, decisions are part of the true causal explanations of actions, because they are none other than (they are strictly identical to) the brain events which cause those actions.

Figure 5.6: Mind–brain identity



The main argument for the thesis of mind–brain identity, then, can be represented as follows:

(1) Our bodily movements are caused by brain events.

(2) Each event in the brain has a sufficient physical cause.

(3) Our decisions are sometimes necessary conditions for some of our movements.

(4) Our decisions sometimes form part of the true causal explanation of some of our movements.

(C) So decisions *are* brain events.

Premises (1) and (2) are intended to rule out classic interactive dualism; premise (3) rejects causal over-determination; and premise (4) rules out epiphenomenalism – thus leaving physicalism as the only remaining possibility.

Premise (1) seems undeniable in the light of modern scientific knowledge. Premise (2), also, seems sufficiently well-supported, given what is known about causal processes in the brain. And premises (3) and (4) form an important part of our common-sense view of ourselves and the world. We believe that mental processes can make a difference to the world. Then the only way

in which we can hold on to this belief, together with the other premises, is to endorse the conclusion – which is the identity-thesis (or at least a limited version of it; see below). The argument as a whole seems rationally convincing in the absence of evidence to the contrary.

Here, as previously, it should be easy to extend the argument to justify the physical nature of *all* mental states, and not just decisions. This is because any mental state can play a part in causing a decision. I sometimes take a decision because of what I see, or what I feel, or what I want, or what I think. In which case these states, too, will form part of the true causal explanation of my action, and the same argument will lead to the conclusion that they, too, are physical.

Given its validity and the strength of its premises, the argument above could reasonably be taken as a proof of the identity-thesis, were it not for the myriad objections which can be raised against that thesis. (Some of these have already been presented in chapter 1:3, in the guise of arguments for the truth of weak dualism.) In sections 3 and 4 below we shall consider a number of them, many of which involve apparent breaches of Leibniz's Law ('identical things share identical properties'). Despite the strength of the argument in its support, the identity-thesis will only be rationally acceptable if we can reply adequately to each (or at least to most) of the objections. (The qualification here is required because it can often be rational to hold onto a theory in the face of some difficulties or 'anomalous data'. Scientists do this all the time.)

2 Ramifications: types, tokens and other minds

In this section we will first clarify the thesis of mind–brain identity, distinguishing between two different versions of it, and relating it to different varieties of reductionism in general. And we will then briefly explore to what extent the truth of that thesis enables us to make progress with the problem of other minds, left over from chapter 1.

2.1 Type versus token identity

There is an important distinction to be drawn between *type*-identity and *token*-identity. The thesis of mind–brain *type*-identity holds that each general type of mental state – for instance, sensations of red, or pains in general – is identical with some general type of brain-state. So whenever a pain is felt it will be identical with a particular instance of some general type of brain-state, the same type of brain-state in each case. The thesis of mind–brain *token*-identity is much weaker. It holds only that each particular instance of pain is identical with some particular brain-state, those brain-states perhaps belonging to distinct kinds. It holds that each particular occurrence of a mental state will be identical with some particular occurrence of a brain-state, but that there may be no general identities between types of mental state and types of brain-state. Note that the arguments for the identity-thesis which we sketched above are indifferent between these two versions of it.

There is some reason to think that the thesis of mind–brain token-identity is the better theory. One argument would be this. We know that there is a considerable degree of plasticity in

the human brain. For instance, although speech is normally controlled from a particular region in the left hemisphere, someone who has had that region damaged (especially when young) can sometimes recover their ability to speak, with practice. So a particular decision to speak may sometimes be identical with an event in one part of the brain, while sometimes it may be identical with an event in quite a different part. Now it doesn't immediately follow from this that the brain events are of different types: this will depend upon what counts as a 'type' of brain-event. But there seems at least no particular reason to *assume* that the events will all be of the same type.

The case can be made even stronger if we recall that many creatures besides human beings can possess mental states. If not only mammals, birds and reptiles, but perhaps also non-biological systems such as robot–computers can possess mental states, then it is obviously false that there will always be the same one type of physical state in existence whenever there exists an instance of a given type of mental state. For the physical control-structures of these creatures will be very different from one another, and from the structure of the human brain. (One aspect of this topic – namely the question of 'artificial intelligence' – will be pursued a bit further in chapter 8.)

Supposing that the thesis of token-identity is the correct theory of mind–brain identity, then there must surely be something more to be said, at the physical level, about what is common to all the different kinds of physical event which are (are identical with) pains. Consider the following analogy. The true theory of clouds is very likely a version of token-identity thesis. For clouds can be made up out of many other kinds of droplet besides water droplets. Thus rain clouds, dust clouds, smoke clouds and clouds of industrial smog are all clouds. Yet there must surely be something common to all these different sorts of collections of particles which explains how they are all *clouds*. And indeed there is: what is in common is a *functional* property of the collections in question, having to do with their weight relative to the surrounding atmosphere, and the way in which they reflect light to give the characteristic appearance of a cloud. So the true theory of clouds is a version of token identity thesis, coupled with an account of the function, or causal role, of the different physical tokens.

Similarly, then, for mental states: the best sort of theory of mental states may be a token-identity thesis, coupled with an account of the causal-roles of the different types of mental state which brings out what all the different tokens have in common. In chapter 7 we will explore just this combination of views, when we come to consider various versions of *functionalism*. All accounts of this sort hold that mental states are to be individuated – or distinguished from one another – in virtue of their distinctive causal roles or functions.

2.2 Reduction versus reductive explanation

Does the thesis of mind–brain identity commit us to *reducing* the mind to the brain? Are we required to say that the human mind is *nothing but* the activity of neurons and groups of

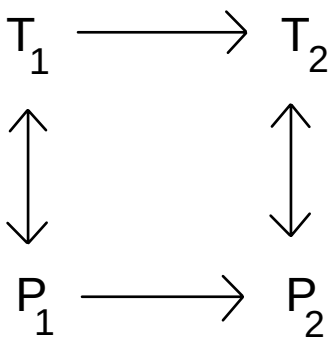
neurons? The answer to these questions is negative, in fact. For a distinction closely related to the one just drawn between type and token identity, is the distinction between *reduction* (of properties), on the one hand, and reductive *explanation* (of tokens), on the other. And we need only be committed to the latter. Let me explain.

Most philosophers and scientists today are physicalists. They believe that all things, events and processes in the natural world are, at bottom, physical things, events and processes. But few are *type*-physicalists, in the sense explained in section 2.1 above. Few believe that higher-level properties in chemistry, biology and psychology, for example, will line up type-for-type with properties in fundamental physics. On the contrary, most believe that the special sciences (chemistry, biology and the rest) are, in a sense, *autonomous* – dealing with laws and properties which cannot be reduced directly to those of physics (or to any other science, indeed).

Admittedly, there have been *some* successful type-reductions of scientific laws and properties. A good example to consider is the reduction of the gas temperature–pressure laws to statistical mechanics. Boyle's gas law states this: PV = kT (pressure × volume = a constant × temperature). So if the volume, *V*, of a gas is kept unchanged, an increase in temperature, *T*, will cause a corresponding increase in pressure, *P*. This law can in fact be derived from statistical mechanics on the assumption that gases are made up of particles in motion, together with the 'bridge principles' that pressure is force per unit area, and that temperature is mean molecular momentum. For as the average momentum of the molecules (the temperature) is increased, so the force per unit area exerted on the surface of the container (the pressure) will also increase, if that surface area remains constant (as it must do if the volume remains unchanged).

The general form of such inter-theoretic reductions can be represented schematically, as in figure 5.7. Here the top line represents a law of the reduced theory, involving the reduced theoretical terms T_1 and T_2, and the bottom line is to be derivable from the laws of the reducing theory, with predicates P_1 and P_2 drawn from some lower-level physical science. The bridge principles ($T_1 \leftrightarrow P_1$ and $T_2 \leftrightarrow P_2$) are then generally thought to license identities between the properties of the reduced and reducing theory.

Figure 5.7: Classical reduction



There exist very few successful inter-theoretic reductions, in fact. The reason lies with the phenomenon of *multiple realizability*. It appears to be quite common for laws in the special sciences (chemistry, biology, neurology, psychology, and so on) to be multiply-realized in lower-level mechanisms. If there are a variety of different physical mechanisms, involving a variety of different physical properties P_i, any one of which is sufficient to realize a property in the special-science law $T_1 \rightarrow T_2$, then it will not be possible to *identify* the special-science property T_1 with any single physical property. This sort of situation is especially likely to arise in the case of biology and psychology, where we know that evolution can come up with a number of different mechanisms to perform the same function. (An example would be the wings of bats, birds and insects, all of which subserve flight, but all of which are structurally very different.) In which case we shouldn't expect to be able to find reductive accounts of psychological properties, including perhaps the properties involved in intentionality or mental representation, on the one hand, nor those involved in consciousness, on the other (these issues will be discussed at some length in chapter 8).

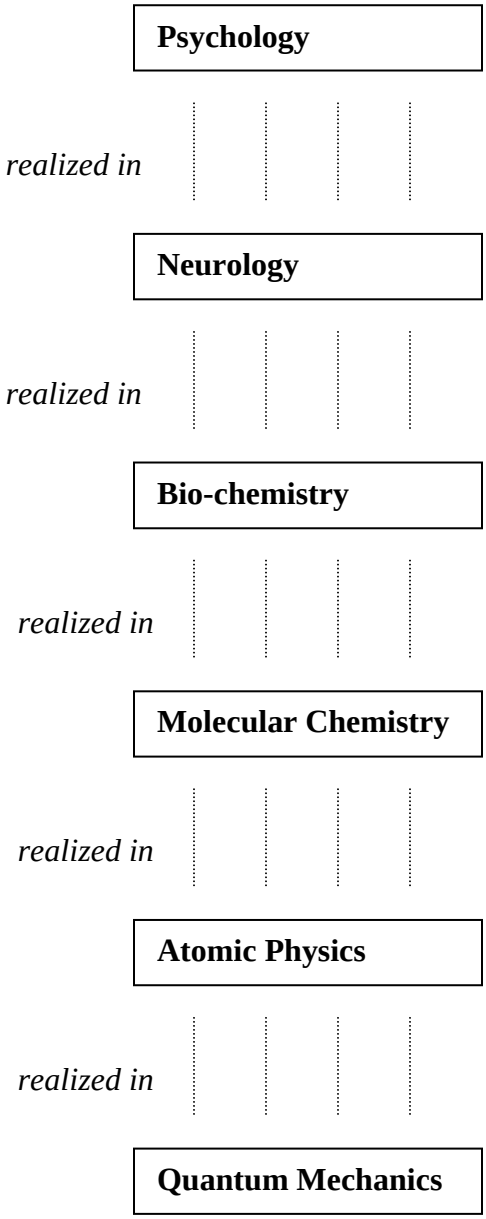
What we *do* regularly find in science, however, is *reductive explanation*. A given higher-level process – in biology, say – is reductively explained when we can show that suitable lower-level event-sequences, happening in accordance with lower-level laws, are sufficient to *realize*, or *constitute*, the higher-level process in question. To put the same point rather differently: a successful reductive explanation shows how a particular instantiation (or type of instantiation) of a higher-level property is constituted by some lower-level property or process. But it does so without reducing the higher-level property as such, since there may be no lower-level process-type which is *always* instantiated whenever the higher-level property is instantiated.

Most physicalists believe in the unity of science, in the sense that they expect all higher-level properties and processes to be reductively explicable in principle. (Such physicalists thus think of the world as ordered into *layers* linked by *realization* relations, somewhat as depicted in figure 5.8.) They think that it must be possible, in the end, to show how any higher-level property or process (of biology, psychology, or whatever) is realized in – or constituted by – some lower-level property or process (and ultimately by processes in fundamental physics). It must be possible to take a particular occurrence of a higher-level property and show how, on at least that occasion, it was constituted by some lower-level physical property or process. But physicalists don't have to say that biology is *nothing but* chemistry, or that chemistry is *nothing but* quantum mechanics.

It is thus possible to be a physicalist about the mind, while at the same time believing in the reality and irreducibility of mental properties, in exactly the sense that it is possible to be a physicalist about wings while at the same time believing in the irreducibility of the property *being a wing*. And while in one sense the mind is *nothing but* the operation of the brain, for a physicalist – since each token mental state will be none other than some token brain state – it may still be the case that if we want to understand the operations of minds in general, we shall unavoidably have to couch our explanations in terms of mental properties; just as if we want to

understand the operations of wings in general, we cannot appeal to the specific physical structures of specific wings.

Figure 5.8: The unity of science and the layered world



2.3 Mind–brain identity and the problem of other minds

Can the thesis of mind–brain identity, if true, provide us with a solution to the problem of other minds? Recall from chapter 1:2 that an argument from analogy to the existence of other minds was obstructed by the claimed uniqueness of my own states of consciousness. But if the identity-thesis is true, then my experiences aren't especially unique, after all. For then they are, in fact,

physical states of the brain, and other people, too, presumably enjoy such states. So an argument from analogy could go through after all, as follows:

(1) I know of the existence of conscious mental states from my own case.

(2) All of my mental states are in fact brain-states.

(3) Other human beings possess brain-states similar to mine.

(C) So other human beings possess mental states similar to mine.

Of course this argument is not strictly valid, since it purports to be a species of inductive argument. But it does appear to be rationally convincing. Moreover premises (1) and (3) are pretty obviously true, while premise (2) merely states the identity-thesis. So if we could know that thesis to be true, the argument as a whole would carry conviction.

If the identity-thesis is true, then there is no longer any problem about knowing that other experiences exist: I can know this to be true on the basis of an argument from analogy. But can I know the particular experiences which people possess on particular occasions? If I see someone injured and groaning, I can know that they possess *some* conscious experience. But can I know that they are aware of the sensation with the distinctive qualitative feel which I describe, in my own case, as ‘pain’? Indeed, is the argument from analogy sufficiently strong to rule out the following sort of possibility: the conscious state which in their case is caused by injury and causes groaning has the qualitative feel which, were I to be aware of it, I should describe as a tickle? (Compare the case of *inverted color experience* which we used in outlining the problem of other minds in chapter 1:1; see figure 1.1.)

Recall the example of the black boxes found on the seashore, which we used in chapter 1:2. Since the boxes all perform the same functions, we are entitled to conclude that they all contain states occupying the same causal roles, namely mediating between a specified input (e.g. a red button being pressed) and a specified output (e.g. a red light flashing). But the states occupying those causal roles may have only that in common: their causal role. In other respects they can be as different as you please.

Similarly, then, in the case of human beings: when I observe an injured person exhibit pain-behavior, I am entitled to conclude that there is some state in them occupying the same causal role occupied, in my own case, by the sensation of pain. And in virtue of the likely truth of the mind–brain identity thesis, I am also entitled to conclude that the state occupying that causal role is very likely a mental as well as a physical one. But it seems left open that it might be quite different in other respects (in particular, in respect of its qualitative feel) from the state which I call ‘pain’.

But what of considerations of simplicity? Isn’t it a great deal simpler to suppose that the same causal roles are occupied by the same feelings in all of us? And aren’t simpler theories in general move reasonable? But in fact the difference in the degree of simplicity here is only marginal. For what really does the work, in our explanations of the behavior of ourselves and others, is the supposition that we all possess states which occupy distinctive causal roles. What explains how you can respond appropriately to the command, ‘Bring me a red flower’, is the fact

that you have learned to discriminate objects on the basis of some-experience-or-other, and to associate with that experience the term ‘red’. Any hypothesis about the particular distinctive feel of your experience seems redundant to the explanation: I don’t in fact need to employ any such hypothesis.

There seems no reason, at this stage, why it should be thought more likely that human beings all have the same feelings occupying the same causal roles, even given the truth of physicalism. For we know already that people differ from one another in all sorts of subtle ways. It is rather as if we had found a set of black boxes which not only share many of their features, but which are each of them, in many ways, unique. None of them look quite alike or has the same physical dimensions, and their responses to any given stimulus (e.g. the pressing of a red button) will often differ quite markedly from one another. Now in cases where we could be confident that all of the boxes had states occupying the same causal roles, could we also be reasonably confident that those states would be similar in other respects? Surely not. In the light of the many differences existing between the boxes, it seems just as likely that the physical mechanisms underlying any given causal role will be subtly different in each case. Equally, then, in the case of human beings: since we already know that they differ from one another in many ways, there is no particular reason to think that they will all have exactly similar brain-states (= feelings) occupying similar causal roles.

All of this is supposing that we only know of the *general* truth of the mind–brain identity-thesis. The situation would surely be different if we also knew some particular identities. If I knew that pains in myself were always identical with a particular type of brain-state, and then discovered that the states which occupy the same causal role in you are also of that type, then I would have to conclude that you, too, feel pain in those circumstances. There can, for the physicalist, be no differences at the level of mentality which don’t reflect differences at the level of the brain. Unfortunately, however, we lack any knowledge of the required identities at this stage.

I conclude that the identity-thesis can only provide us with the general knowledge that other people possess mental states of some sort, occupying similar causal roles to our own. While this is an advance, it is not fully satisfying. For our common-sense view is that we can often know what other people are feeling on particular occasions. At this stage, however, the Cartesian conception of the meaning of mental-state terms (outlined and defended in chapter 1:4) remains unchallenged. And neither have we enquired whether it is possible to seek a scientific understanding of the nature of conscious feelings. These tasks will be taken up in chapters 7 and 8 respectively.

3 Difficulties for mind–brain identity

In section 1 we presented an argument in support of the identity-thesis which would have been convincing if considered purely on its own terms. Over the next two sections we will consider all of the main objections which have been raised against that thesis, beginning with some of the

less serious ones. They will get more serious as we go along.

3.1 Certainty

Our first objection derives from Descartes, who deployed a similar argument in support of strong dualism. It runs as follows:

(1) I may be completely certain of my own experiences, when I have them.

(2) I cannot have the same degree of certainty about the existence of any physical state, including my own brain-states.

(C) So (by Leibniz's Law) my conscious experiences aren't in fact identical to brain-states.

Although both the premises in this argument are true, the argument itself commits a fallacy, and is invalid. For as we noted in chapter 3:1, Leibniz's Law only operates in contexts which aren't *intentional*. And it is obvious that the context created by the phrase 'X is certain that...' is an intentional one.

For example, the police may be certain that Mr Hyde is the murderer, while they have no inkling that Dr Jekyll is the murderer, despite the fact that Jekyll and Hyde are one and the same man. And Oedipus may be certain that Jocasta loves him without believing that his mother loves him, despite the fact that Jocasta is his mother. So from the fact that I have complete certainty about my own conscious states without having certainty about my own brain-states, it doesn't follow that my conscious states aren't brain-states. For just as one and the same woman may be presented to Oedipus in two different guises – as Jocasta, and as his mother – so perhaps one and the same brain-state may be presented to me under two different aspects: in a third-person way (*as* a brain-state), and via the qualitative feel of what it is like to be *in* that state.

3.2 Privacy

This second argument is a variation on the first. It runs as follows:

(1) Conscious states are private to the person who has them.

(2) Brain-states aren't private: like any other kind of physical state, they form part of the public realm.

(C) So (by Leibniz's Law) conscious states aren't in fact identical to brain-states.

The term 'private', here, is ambiguous, however: something can be private in respect of *knowledge* (only I can know of it), or it can be private in respect of *ownership* (only I can possess it). Taken in the first way, the argument is the same as that in section 3.1 above, and commits the same fallacy. But taken in the second way, premise (2) is false.

It is true that only I can 'own' my conscious states: no one else can feel my pain, or think my thought. But it is equally true in this sense that only I can own my brain-states. For other people can't possess my brain-state either. Any brain-state which they have will be, necessarily, their own brain-state, not mine. Conscious states are certainly not unique in respect of privacy of ownership. The same is true of blushes and sneezes as well as brain-states: no one else can blush

my blush or sneeze my sneeze. Indeed it seems that in general the identity-conditions for states and events are tied to the identities of the subjects who possess them.

3.3 Value

A thought can be wicked. A desire can be admirable. But no brain-state can be either wicked or admirable. We may therefore argue as follows:

(1) Mental states are subject to *norms*: they can be good or bad, wicked or morally admirable.

(2) No purely physical states are subject to norms: no brain-state can be either wicked or morally admirable.

(C) So (by Leibniz's Law) mental states aren't identical to brain-states.

We might immediately be inclined to quarrel with premise (2). For can a particular stabbing not be wicked? And what is a stabbing if not a physical event? But it may be replied that it is not the stabbing itself – considered merely as a physical event – which is wicked, but rather the intention behind it. And in general, physical states and events are only subject to moral norms if they are intended (or at least foreseen). This is because norms imply *control* – or as it is sometimes said, 'Ought implies can'.

This reply is sufficient to save the truth of premise (2), but at the cost of revealing the same fallacy in the argument as was involved in the argument from certainty. For if it is only things which are intended (or foreseen) which can be wicked, then the context created by '. . . is wicked' will be an intentional one. For example, if the fact that it is wicked of Mary to have a particular desire implies that she either intentionally adopted that desire, or at least foresaw that she would continue to possess it if she took no steps to eradicate it, then it is no objection to the identity-thesis that it is, on the other hand, not wicked of her to be in such-and-such a brain-state. For you can foresee *e* (that you will marry Jocasta) without foreseeing *f* (that you will marry your mother), even though *e* is identical to *f*. And certainly Mary neither intended nor foresaw that she should be in that particular brain-state.

3.4 Color

An after-image can be green. A pain can be sharp and piercing. But it is hardly likely that any brain-state will be either green or sharp and piercing. There will then be many arguments which take the following sort of form:

(1) I am experiencing a fading green after-image.

(2) No brain-states are green.

(C) So (by Leibniz's Law) my after-image isn't identical to any brain-state.

One mistake in this argument is that it treats my after-image as though it were a particular individual thing, having greenness as a property. Now, it is true that the sentence, 'I experience a green after-image' has the same grammatical form as the sentence, 'I pick up a green book', which creates the impression that greenness is a property of the after-image in just the same way

that it is a property of the book. But this impression is misleading. For the book in question might have had some other color, while remaining the numerically-same book. But can we make any sense of the idea that the very same after-image which I now have might have been red?

What is true, of course, is that I might now have been experiencing a red after-image rather than a green one. But can we make sense of the idea that it might have been this very same after-image (which happens to be green) which would then have been red? I suggest not. Rather, the greenness is essential to the identity of that particular after-image. And this is because it is a mistake to treat the fading after-image as if it were a kind of object or individual thing. It is rather an event, or happening. And the greenness is in fact *part* of the event of experiencing-a-fading-green-after-image, as opposed to being a property *of* it.

This reply on its own isn't sufficient to rebut the argument. For it seems certain that greenness is *not* a part of the event of undergoing-such-and-such-a-change-in-brain-state. So how can that event be identical with the event of experiencing-a-fading-green-after-image, if greenness really is part of the latter event? For must not identical events have identical parts? For example, if the battle of Waterloo is (is identical with) the battle which lost Napoleon the war, then if a particular cavalry charge is part of the battle of Waterloo, it must also be part of the battle which lost Napoleon the war. For they are the very same battle (the very same event).

The second mistake in the argument is to think that an experience can be green – or that a mental event can have greenness as one of its parts – in anything like the same sense that a physical object can be green, or that a physical event can have greenness as one of its parts. Experiences aren't literally green in the way that grass is literally green. Rather, the 'greenness' of my green after-image consists in my having an experience which is *like* the experience of *seeing* a green patch (and it is the patch which is green, not the experience). Then if having a green after-image is like seeing a green patch, and if the possession of that after-image is identical with some brain-state, it will indeed follow that the brain-state, too, is like the state one is in when one sees a green patch. But there is no difficulty about this. For the latter state will also be identical to some brain-state. And it seems entirely plausible that there should be some resemblance between the two brain-states.

The point can be put like this: green after-images are experiences *of* green, rather than things (or events) which *are* (or which contain) green. A green after-image is, as it were, an event of being-under-the-impression-that-one-is-seeing-a-green-patch. If this is so, then the after-image can be identical with a brain-state without breaching Leibniz's Law, so long as the brain-state, too, can in this sense be 'of' green. If green after-images *represent* green, rather than literally *being* green, then there will be no difficulty here so long as it is possible for brain-states to represent things. This question will be pursued briefly in section 3.8 below, and then again at greater length in chapter 8.

3.5 Felt quality

This objection arises out of the last one. Even if after-images aren't literally colored, still they do

literally have phenomenal (felt) characteristics, which we describe by means of the color-terms (‘sensation of red’, ‘experience of green’, etc.). There is the qualitative ‘feel’ which is common to the experiences of having a green after-image and seeing green grass, for example. But can any brain-state have a distinctive feel? For example, think of a particular brain event – say one group of brain cells firing off impulses to another group – and ask yourself, ‘What is the distinctive feel of this event?’ It is hard even to get a grip on the question. How can any brain-event *feel like* a sensation of red, for example? Surely only another sensation can, in the required sense, have the *feel* a sensation of red.

The argument sketched above may be summarized as follows:

(1) All experiences have distinctive felt qualities.

(2) Brain-states don’t have distinctive felt qualities.

(C) So (by Leibniz’s Law) experiences aren’t in fact identical with brain-states.

The problem, here, is to know that premise (2) is true. For if the thesis of mind–brain identity is correct, the felt quality of an experience *is* some brain-state or property of a brain-state. And the difficulty we have in seeing that this is so may relate entirely to the different perspectives which we take on this one-and-the-same state. We can think about that state in a third-person way, in terms of spatial extent, electrical potentials, chemical reactions and so on; or we can think of that state in a first-person way, in terms of the way that it feels to a subject who is *in* that state.

There is nothing here, as yet, to convince us that the felt properties of experiences aren’t properties of the brain. Yet it is one thing, of course, to accept that these two sorts of properties may really be identical (or, in the light of the arguments of section 1 above, to accept that they *are* identical), and it is quite another thing to understand *how* this can be so. Our task at this point is just to show that there isn’t any good reason *not* to accept the arguments of section 1, and so no good reason *not* to believe in physicalism. In chapter 8 we will return to consider whether it is possible to give some satisfying *explanation* of how felt qualities can be physically constituted.

3.6 The explanatory gap

This objection again arises quite naturally out of the last one. For some have claimed that there is an unbridgeable explanatory gap between all physical facts, on the one hand, and the facts of felt consciousness, on the other. (This idea was first developed by the American philosopher Joseph Levine.) No matter how detailed a story I am given about the operations of the brain, I shall always remain capable of thinking, ‘Surely all *that* might be true while *this* sort of feeling was different or absent’. (It is this thought which suggests the possibility of zombies, indeed.) And some have claimed that the truth of physicalism *requires* that conscious states should be reductively explicable in physical terms. They can then argue as follows:

(1) If the thesis of mind–brain identity were true, then it would have to be possible to *explain* the felt qualities of our experiences in physical terms.

(2) No such explanation is possible: there is an explanatory gap between physical facts and feeling facts.

(C) So the thesis of mind–brain identity isn't true.

This argument is valid. But we can explain why reductive explanation of felt qualities is impossible, however (at least when those qualities are conceptualized in a certain way), consistent with the truth of mind–brain identity – in effect, denying premise (1). We just have to recall that we may possess some *purely recognitional* concepts of experience, as was argued when we were defending the Cartesian conception of the meaning of mental-state terms in chapter 1:4. For if so, then no story about physical events and causal roles will be capable of engaging with these concepts, leading us to think, 'Ah, in those circumstances *this* state would have to be present'. On the contrary, nothing will be capable of evoking an application of one of these concepts expect the actual presentation of the appropriate sort of felt quality. But for all that, the quality recognized may actually be a physical one, just as the thesis of mind–brain identity affirms.

We will return to consider the challenge of providing a reductive explanation of the felt qualities of consciousness in chapter 8. For the moment, we can note that although the explanatory gap might show something distinctive about our *concepts* for felt qualities of various types (namely, that they are purely recognitional concepts, lacking conceptual connections with concepts of other sorts), the existence of such a gap shows nothing about the non-physical status of the felt qualities themselves.

3.7 Complete knowledge

Someone could know all physical and functional facts about the brain without knowing what the different experiences feel like. In order to see this, consider the example of color-deprived Mary, first introduced into the now-extensive literature by the Australian philosopher Frank Jackson.

We are to imagine the case of Mary, who has lived all her life in a black-and-white room. At the point where we take up the story, Mary has never had any experience of color; but, we may suppose, there is nothing wrong with her visual system – she still has the *capacity* for color vision. Now, Mary is also a scientist, living in an era much more scientifically advanced than ours. So Mary may be supposed to know *all there is to know* about the physics, physiology, and functional organization of color vision. She knows exactly what takes place in someone's brain when they experience red, for example, and has full understanding of the behavior of the physical systems involved. So she knows all the objective, scientific, facts about color vision. But there is one thing she *doesn't* know, surely, and that is what an experience of red *is like*. And on being released from her black-and-white room there is something new she will *learn* when she experiences red for the first time. Since knowledge of all the physical and functional facts doesn't give Mary knowledge of *all* the facts, Jackson argues, then there are some facts – namely, facts about subjective experiences and feelings – which aren't physical or functional facts, and which cannot be explicable in terms of physical or functional facts, either.

The thought is: if there is information about feelings which could not be conveyed by any amount of information about the brain, then feelings cannot themselves be brain-states. The

argument is as follows:

(1) Even complete knowledge of physical states wouldn't give someone the knowledge of what an experience *feels like*.

(2) But if experiences *were* physical states, then complete knowledge of the physical states *would* imply complete knowledge of experiences, including knowing what they *feel like*.

(C) So experiences aren't physical states.

Although this argument, like the argument from certainty in section 3.1, involves an intentional term (the context created by the phrase 'X knows that . . .' is an intentional one) it doesn't seem to commit the same fallacy. This is because the premises speak of *complete* knowledge, knowledge from all points of view. Oedipus certainly couldn't have complete knowledge of Jocasta without knowing that she is his mother. (So if he does know everything about her, but doesn't know that she is his mother, then she *isn't* his mother.)

All the same, the argument is fallacious. In the sense of 'complete knowledge' in which premise (1) can plausibly be thought to be true, premise (2) is false. (And conversely, in the sense of 'complete knowledge' required for premise (2) to be true, premise (1) is false.) In order to see this, notice that there are two ways of counting items of knowledge – either in terms of the worldly facts known about, or in terms of the beliefs or sentences which represent those facts. If Oedipus knows that Jocasta is beautiful and knows that his mother is beautiful, then we can *either* say that there is just one item of knowledge involved (the knowledge, namely, of the beauty of a particular woman), *or* we can say that there are two items of knowledge involved (one for each of his distinct beliefs).

Notice, now, that for premise (1) to be plausible, items of knowledge have to be counted in the first of these ways. For if items of knowledge were counted in terms the distinct ways of thinking of a subject-matter, then it would be *impossible* to have complete knowledge of *anything*! For any one fact can always be represented in *infinitely* many distinct ways. For example, the fact that a particular brain-cell fires can be represented by the sentence, 'The brain-cell exactly 1.11111 millimeters below this point in the skull is firing', or by the sentence, 'The brain-cell exactly such-and-such a distance below this point on the ceiling is firing', or by the sentence, 'The brain-cell exactly such-and-such a distance from this point in the roof is firing', and so on, and so on.

Having 'complete knowledge' of one's brain-states surely couldn't require one to know all facts about the brain under all possible modes of presentation, or all ways of thinking of those facts. (And if it did, then premise (1) would just beg the question against the identity-theorist. For amongst all these ways of thinking of a brain-state, according to the physicalist, will be the way of thinking given in terms of what it feels like to be in that state.)

But now, if 'complete knowledge' is read as meaning 'knowing all facts in terms of *some* (not all) ways of thinking of them', then premise (2) is false. From the claim that someone knows all of a certain range of facts represented in *some* way, it doesn't follow that they know

those same facts represented in *all* ways. And if the identity-thesis is correct, then amongst these other ways of thinking will be first-person ways of thinking, grounded in the way experiences (= brain states) feel to subjects who have them.

So we can allow that there are some things that color-deprived Mary won't know about color vision, even if she knows all facts about the brain. For there are some concepts which you can only have if you have undergone certain kinds of experience – namely, recognitional concepts of the *feels* of those experiences. And so there are some thoughts which will be unavailable for Mary to think – namely, thoughts employing those recognitional concepts. And then there will be some thoughts which Mary can't know to be true, either. But if the identity-thesis is correct, then some of these thoughts will be *about* facts which Mary *does* know to be true. What Mary would gain, if she could acquire color vision, would be some new concepts and ways of thinking of the very same brain-events which she would already have had scientific knowledge of.

3.8 Intentionality

Recall from chapter 1:3 the claim that mental states are unique in being *intentional* (i.e. representational). Our argument was as follows:

(1) Some mental states are intentional, or representational, states.

(2) No merely physical state (e.g. of the brain) can be intentional in its own right.

(C) So (by Leibniz's Law) some mental states aren't identical with brain-states.

This argument is valid, and premise (1) is obviously true. So everything turns on the acceptability of premise (2).

There is a general philosophical problem about representation. Much of the philosophy of mind over the last twenty years has been concerned with the question: how is it possible for anything to represent – or be about – anything else? It is by no means easy to *understand* how an arrangement of cellular connections could represent anything. This is a topic we will return to in chapter 8. Here we will do just enough to show that there need be no convincing reason to believe that premise (2) is true.

Perhaps the claim that physical states can be representations-in-their-own-right cannot be made *entirely* convincing in the absence of a solution to the general problem of representation. But we can at least get an inkling of how intentionality can be embodied in a physical system in advance of a solution to that problem. For by looking at systems which are, manifestly, purely physical – namely, computers and computer-controlled machines – we can begin to see how they can display some of the distinctive features of intentionality. If we can see the beginnings of intentionality embodied in a physical system such as a computer, then there is no reason in principle why full-blown intentionality (beliefs, desires and the rest) shouldn't be embodied in the biological computer which is the human brain.

One distinctive feature of intentional states is that they represent things in one way rather than another. For a crude analogue of this, imagine a computer linked to a video-camera and

mechanical arm. The computer is programmed to scan the input from the camera, and to grab with its arm any yellow object. In order for the grabbing-operation to be successful it must also be able to interpret from the input the shapes, sizes and spatial positions of those objects. But the computer doesn't select objects on the basis of their shape or size, but only on the basis of their color.

Now suppose that the only yellow objects which are ever presented to the machine are in fact lemons, its purpose being to select lemons from a passing array of fruit. Of course lemons do also have a characteristic shape, but the computer is indifferent with respect to shape. It initiates a grabbing motion only in response to the yellow color. Then there is almost a sense in which the machine might be said to desire the yellow objects which it grabs, but not the lemon-shaped objects, even though the yellow objects are all lemon-shaped.

I don't want to say that such a machine would literally have a desire for yellow objects, of course. Although quite what is missing here, which would be present in the case of a genuine desire, is not easy to see. Perhaps (as we suggested in chapter 2:4) we can only make sense of something having a particular desire against a wider background, a network of other desires and beliefs. Or perhaps only a being which is alive, which has needs (and which may consequently be said to have 'a good') can have desires. Some of these possibilities will be explored in chapter 8. The important point for our purposes here is that we have found an analogue for the intentionality of desire in the concept of 'differential response'. Just as Oedipus will respond differently to one and the same woman presented to him now as Jocasta, now as his mother; so the machine will respond differently to one and the same bit of fruit presented to it now as a yellow thing (its shape being obscured), now as a lemon-shaped thing (its color being obscured).

The other distinctive feature of intentional states, is that they can be directed at non-existent objects. And it is apt to seem unintelligible how any physical system could do this. Here again my strategy is to construct, by way of reply, a crude physical analogue for this aspect of the intentionality of the mind. Thus consider the sort of behavior which might be displayed by a Cruise missile. It is programmed to take photographs of the terrain beneath it at various points along its route, to scan those photographs for landmarks in order to check its position, and adjust its direction accordingly. Now suppose that as a result of an error, it is programmed to find a distinctively-shaped lake at a particular point on its route, but that no such lake exists. As a result, the missile circles round and around the area, until finally it runs out of fuel and crashes. Here we might almost say, 'The missile was searching for a lake which didn't exist.'

Note that the intentionality displayed in the Cruise missile's 'desire' isn't merely derivative from the thoughts and intentions of the computer-programmer. True enough, that 'desire' was caused to exist by the programmer. But its intentionality – its directedness on a non-existent object – is actually displayed in the behavior of the missile itself. For it has entered a cycle of behavior which we know will only be terminated if it succeeds in photographing a lake with a particular distinctive shape; but we also know that no such lake exists.

I tentatively conclude that there is no reason in principle why a merely physical system

shouldn't display the various features characteristic of intentional states. So we have been given no reason for supposing that beliefs and desires aren't themselves physical (brain) states. The point is, arguing that intentional states *can't* be physical states is one thing (and that argument can be seen to fail); achieving a detailed *understanding* of *how* intentional states can be physical ones is quite another (and that is something I don't pretend to have provided, either here or in chapter 8).

3.9 Free will

Many people believe that humans have free will, in the sense that they can make decisions which aren't determined by prior causes. But it looks certain, in contrast, that each brain-event will be determined by prior causes. We may therefore argue thus:

(1) The decisions people make can be free, not determined by prior causes.
(2) All brain events are determined by prior causes in the brain or central nervous system.
(C) So decisions (or at least the free ones) aren't identical with brain events.

This argument is valid. The real question is whether the two premises can be adequately supported. This is a large topic, which we must set aside for the moment. We will return to it at some length in chapter 8. (And remember that we don't necessarily have to be able to reply to *every* objection to the identity-thesis in order for the latter to be rationally acceptable. Some of these objections can be left as *anomalies* which we don't presently know how to solve.)

3.10 Spatial position

Recall from chapter 1:3 the following argument:

(1) All brain-states must occupy some particular position in space.
(2) It is nonsense (meaningless or self-contradictory) to attribute any particular spatial position to a mental state.
(C) So (by Leibniz's Law) conscious states can't be identical with brain-states.

This argument is valid. Premise (1) is obviously true. So everything depends upon the truth of premise (2). I shall not waste time quibbling that some mental states (e.g. pains) are apparently attributed spatial positions. For the identity-thesis extends to all mental states without exception. And in any case it is unlikely that some mental states are identical with brain-states while some are not. I shall focus on the hardest case for the identity-theorist; namely, thoughts.

Identity-theorists might be tempted to respond to the above argument by conceding that our ordinary concept of thought makes attributions of spatial position to them nonsensical, but by rejecting that ordinary conception as mistaken. They may insist that every thought (properly conceived of) does in fact have a place, namely the place of its identical brain-state. But for them to take this line might be a mistake. For then the thesis of mind–brain identity would no longer represent a empirical discovery, but would be something which we have stipulated as true through a change of meaning.

We might be entitled to reply to the identity-theorist as follows, indeed. 'If you mean the

word “thought” as we usually do, then your thesis is false; indeed necessarily false. But if, on the other hand, you wish to give the word a different meaning for your own special purposes, then you are perfectly entitled to do so. But don’t pretend that you have made a momentous discovery, or that you are saying anything which conflicts with what the weak dualist believes. All you have done is to give a new definition.’ (It is rather as if someone were to give new definitions of the words ‘red’ and ‘green’, in such a way that it then makes sense to ascribe those words to numbers; and were then to announce, as if they had discovered something terribly important, ‘Contrary to what has always been believed, every even number is red and every odd one green.’)

A more promising strategy for the identity-theorist is to suggest that we have again been misled by the grammatical form of phrases like ‘my thought of my mother’ into conceiving of a thought as if it were a special kind of individual thing or object. For obviously, if a thought were really a physical object like a grain of sand or a brain-cell, then it would have to occupy some precise position in space. So perhaps we need to be reminded that a thought is an action, and an action is a species of event (a ‘happening’). And then the general question becomes: what are the criteria for attributing spatial positions to events?

Often the place of an event can be pinned down no more precisely than the place of the subject of that event. And in such cases requests for more precise specifications will seem nonsensical. Thus the place of the event of Mary-growing-older is wherever Mary is. And the question, ‘Is the event of Mary-growing-older taking place two inches behind her right eye?’ seems just as nonsensical as the parallel question about the event of Mary-thinking-of-her-mother. Yet, for all that, the process of ageing is a purely physical one.

Now we only need to be reminded of these facts to realize that we do in fact attribute spatial positions to thoughts; namely: whenever we say where the *thinker* of that thought is. And the fact that it sounds nonsensical to request more detailed specifications of the spatial positions of thoughts needn’t show that thoughts themselves are non-physical. Thinking, like ageing, may be a physical process whose subject is the whole human being.

The position of an event isn’t always simply the position of its subject, however. An event can also take place in *part* of its subject. Thus the position of Mary’s-left-big-toe-turning-blue isn’t simply wherever Mary is. It is, more precisely, wherever her left foot is. For if Mary is lying on a river-bank with her left foot in the cold water, then the event takes place in the water; whereas Mary herself is *not* (or is only partly) in the water.

It seems likely that the physical event which is, according to the identity-theorist, the event of Mary-thinking-of-her-mother, takes place in some particular region of her brain. So we still have a problem, if the closest we can get to the spatial position of Mary’s thought is the spatial position of Mary. If the brain-event can be two inches behind her right eye, whilst it is incorrect to describe the thought of her mother as occurring two inches behind her right eye, then the thought and the brain-event cannot be identical.

The correct way for an identity-theorist to respond, is by denying that it is nonsensical to

ascribe precise spatial positions to thoughts. The only real evidence which the weak dualist has for this claim, is that most of us would be left gaping, our minds completely blank, if asked whether or not Mary's thought is two inches behind her right eye. But this doesn't show that the question is literally meaningless, nor that it is self-contradictory. It only shows, firstly, that it is not the sort of question which itself points you in the direction in which you have to look for an answer. (Contrast: 'Where is the dam cracking?' It is part of the ordinary notion of a crack, that in order to find them you have to search in specific locations.) And secondly, that we can have no idea where to look for an answer until we have acquired some further information.

Suppose you were asked, 'In what specific region of her body is the event of Mary-catching-a-cold taking place?' This, too, would have a tendency to make your mind go blank, partly because it isn't part of the ordinary concept of a cold that in order to establish whether someone has a cold you have to search in specific locations within the body. But on reflection you may realize that what you are really being asked is: 'In virtue of changes in what parts of Mary's body is it becoming true that she has a cold?' This is a question which you can understand, at least. But if you know nothing of viruses, or of medicine generally, you may not even know what *sorts* of things would be relevant to the discovery of the answer. Yet when you are told that colds are viruses, and that viruses enter the body at specific locations, then you *do* know what would constitute an answer. Note, moreover, that it would be hardly very plausible to say that the term 'a cold' had changed its meaning for you when you acquired this information.

The question about the specific location of Mary's act of thinking about her mother is essentially similar. The first step in dispelling the puzzlement which it causes is to realize that what we are in fact being asked is: 'In virtue of changes in what specific region of Mary is it becoming true that she is thinking of her mother?' The next step is to learn that each mental event is identical with some brain-event. Then we know that in order to answer the question, we should first need to discover *which* brain-event is identical with Mary's thought, and then discover *where* that brain-event is occurring. Yet we don't need to regard our acceptance of the identity-thesis as altering the meaning of the term 'thought', any more than our acceptance that colds are viruses alters the meaning of the term 'a cold'. We may thus reasonably deny the claim made in premise (2) of the argument above.

4 The necessity of identity

In this section we shall deal with a particularly important difficulty for the thesis of mind–brain identity, which turns on the claim that a statement of identity, if true, is true necessarily: it is a truth about all possible worlds. For then if, as seems plausible, the thesis of mind–brain identity is merely contingent (i.e. *not* a truth about all possible worlds), it will follow that it is not true at all. This argument was first developed by the American philosopher Saul Kripke. It closely parallels the one we considered (and dismissed) in chapters 2:1 and 3:4 in support of strong dualism.

4.1 The argument

Thus far in this chapter we have tacitly assumed that the thesis of mind–brain identity is not only empirically grounded but contingent. We have assumed that, although it may be true in the actual world, there are other possible worlds in which it is false. Now what is certainly the case is that it is not a *conceptual* truth: it is not true in virtue of the meanings of the terms involved. But as we saw in chapter 3:4, some necessary truths aren't conceptual truths. Some truths are truths about all possible worlds without being true in virtue of meaning. For example, consider once again the identity between Jekyll and Hyde (supposing them to have been a real historical character). The truth of, 'Jekyll is Hyde' is certainly not merely a matter of meaning; for the police had to discover it by empirical investigation. But it is, for all that, a necessary truth. Since it is in fact true that Jekyll is identical with Hyde, things could not have been otherwise. For if Jekyll is Hyde, then there is only one thing involved rather than two. It is not as if there were two logically distinct things, which happen to be related to one another in a particular way in the actual world ('being identical with one another'), but which could exist unrelated in some other possible world. Rather, there is only one thing, which must remain identical with itself in all possible worlds in which it occurs.

In general, where we have a true identity-statement which involves two names for the same thing, we cannot say, 'This thing is identical with that thing in the actual world, but there are other possible worlds in which they aren't identical.' For if the identity-statement is true, then there is really no 'this' and 'that'. There aren't two things in question, but only one. And it is impossible to conceive of a world in which that thing isn't identical with itself.

Confining ourselves just to the case of pain, then, opponents of the mind–brain identity thesis can now argue as follows:

(1) If each pain is identical with some brain-state, then the things which are, in this world, the pains, are identical with those brain-states in all possible worlds in which they exist. (*Necessity of identity.*)

(2) Each pain in this world is, in some other possible worlds in which it exists, *not* identical with any of those brain-states.

(C) So it isn't the case that each pain is identical with some brain-state, even in the actual world.

Since this argument is valid, and since premise (1) is true, the identity-thesis will have been refuted if we can establish premise (2).

Now premise (2) can in fact be made to seem extremely plausible. I can imagine a world in which the very same pain which I feel at the moment isn't identical with any brain-state. I can, for example, conceive of turning into a pillar of salt (like Lot's wife in the bible story) while the pain goes on; or I can imaging being transformed gradually into a partial humanoid robot, with hard-ware rather than wet-ware encased in the relevant portions of my skull, again while my pain exists unchanged; and so on. Alternatively, in connection with any particular candidate brain-state, I can imagine my current pain existing while *that* brain-state doesn't occur (even

though some other one does). Then since this exercise can be repeated for all other candidate brain-states, and all other pains, we have apparently done enough to establish premise (2).

We have arrived at a powerful-looking argument against the identity-thesis. On the one hand it seems that the very same mental states which I now enjoy might continue to exist, or might have existed, in the absence of any of the relevant brain-activity. But on the other hand it seems that if these mental states are in fact brain-states, then they will have to remain brain-states (i.e. remaining identical with themselves) in all possible worlds in which they occur. It therefore seems that the identity-thesis must be false.

Notice that the argument here would work equally well if premise (2) were replaced with premise (2*), claiming that it is possible to have the relevant brain states without the presence of pain, thus:

(2*) Each of the brain states which is correlated with the presence of pain in this world, can occur in other possible worlds *without* the existence of any pain.

What this premise tells us, in effect, is that philosophical zombies are possible. And since it is possible for me to be physically just as I am without feeling any pain (a zombie), my pains cannot be identical with any physical states. For if they *were* identical, then (by the necessity of identity) they would have to exist whenever the relevant brain states exist, and zombies would be *im*possible.

4.2 Criticism of the argument

I shall now defend the claim that premise (1) of the argument above is only true in respect of *metaphysical* (as opposed to *conceptual*) necessity; but that the most that we have reason to believe with respect to premise (2) (and also (2*)) is that it is *conceptually* possible that the feel of pain in question isn't identical to the candidate neural event in question. So the argument is, after all, invalid. In so far as we have sufficient reason to believe its two premises, it commits a fallacy of equivocation (that is, a fallacy of ambiguity, or a shift in meaning). For premise (1) is a metaphysical truth, whereas premise (2) (and also (2*)) is a merely conceptual one.

Recall that the moral of the story of Jekyll and Hyde, discussed in chapter 3:4, was that not all necessities and possibilities are conceptual ones. Something can be conceptually possible (conceivable) while being metaphysically *im*possible; and something can be metaphysically necessary which isn't conceptually so. Indeed, precisely this sort of situation will arise whenever we conceive of what is in fact one and the same thing or event in a number of distinct ways. Then with the distinction between conceptual and metaphysical necessities firmly in place, we need to enquire after the status of the premises of the above argument. In particular, is the possibility that the feel of pain might exist, or might have existed, independently of its associated brain-state a genuinely metaphysical possibility, or is it merely conceptual?

All of the kinds of data which seem to establish the truth of premise (2) have to do with *conceivability experiments*, in fact. Thus, I can *conceive of* turning into a pillar of salt while the pain continues; or I can *imagine* being transformed into a partial humanoid robot while my pain

exists unchanged; and so on. And the most that such thought experiments can establish is that it is *conceptually* possible that the pain in question isn't identical with the given neural event. But then that isn't enough to generate an argument against physicalism, any more than the conceptual possibility that Jekyll isn't Hyde is any argument against their actual identity.

Similarly, the argument giving rise to premise (2*) is also a mere conceivability experiment. We can, indeed, *conceive of* me existing physically exactly as I am now, while nevertheless lacking any experienced pains. And so it is, indeed, conceptually possible for zombies to exist. But it doesn't follow that this is also metaphysically possible – it doesn't follow that there is really any world in which someone physically indistinguishable from me lacks pain. Indeed, if the thesis of mind-brain identity is true, then such a thing *isn't* possible, just as it isn't really possible to have a world in which Jekyll exists but Hyde doesn't.

If Jekyll *is* Hyde, then Jekyll will remain Hyde in all possible worlds in which they exist – the identity, if true, is metaphysically necessary, despite the fact that it is easy to *conceive of* Jekyll existing while Hyde does not. Similarly, then, in respect of mind and brain. If this pain *is* a particular neural state, then this pain will remain that neural state in all possible worlds in which it exists, despite the fact that we can *conceive of* the pain continuing, or of it having existed, in the absence of that neural state; and despite the fact that we can *conceive of* zombies. There is no good argument here against the thesis of mind–brain identity.

Conclusion

In section 1 of this chapter we presented arguments for thinking it likely that all mental states are identical with brain-states. In section 2 we saw that this thesis is most plausible when confined to *tokens* (as opposed to *types*) of mental state, and we saw that the thesis provides a partial solution to the problem of other minds. Then in the sections following we have replied to all (or at least most of) the various objections to the identity-thesis. Since there is good reason to believe the identity-thesis to be true, and no good reason (as yet) to believe it false, the case for that thesis is rationally convincing. We should therefore embrace the thesis of mind–brain identity, and declare ourselves to be physicalists about the human mind.

Questions for discussion

1. How strong are the arguments for thinking that all mental states are identical with brain states? Is this the only way in which we can believe that the mind makes a causal difference?
2. If mental states were merely non-physical a-causal *epiphenomena* of the brain, then how could we ever tell?
3. Must physicalists about the mind be committed to saying that the mind is *nothing but* the brain?
4. Is there something about color experience which Mary couldn't know, while locked in her black-and-white room, which she would learn as soon as she comes out? If so, what (if anything) does this show?

5. Can you conceive of turning into a pillar of salt (like Lot's wife) while your headache continues? If so, what consequences would this have for the identity-thesis?

Further reading

David Armstrong, *A Materialist Theory of the Mind* (Routledge, 1968).

Donald Davidson, 'Mental events', in L. Foster and J. Swanson (eds.), *Experience and Theory* (Duckworth, 1970). Reprinted in N. Block (ed.), *Readings in the Philosophy of Psychology*, vol. 1 (Methuen, 1980); and in Davidson's *Actions and Events* (Oxford University Press, 1980).

Frank Jackson, 'Epiphenomenal qualia', *Philosophical Quarterly*, vol. 32 (1982). Reprinted in W. Lycan (ed.), *Mind and Cognition* (Blackwell, 1990).

Frank Jackson, 'What Mary didn't know', *Journal of Philosophy*, vol. 83 (1986). Reprinted in N. Block *et al*. (eds.), *The Nature of Consciousness* (MIT Press, 1997); and in D. Rosenthal (ed.), *The Nature of the Mind* (Oxford University Press, 1991).

Saul Kripke, *Naming and Necessity* (Blackwell, 1980), lecture 3.

Joseph Levine, 'Materialism and qualia: the explanatory gap', *Pacific Philosophical Quarterly*, vol. 64 (1983).

David Lewis, 'An argument for the identity theory', *Journal of Philosophy*, vol. 63 (1966). Reprinted in Lewis' *Philosophical Papers*, vol. 1 (Oxford University Press, 1983); and in D. Rosenthal (ed.), *Materialism and the Mind-Body Problem* (Prentice Hall, 1979).

Thomas Nagel, 'What is it like to be a bat?', *Philosophical Review*, vol. 83 (1974). Reprinted in N. Block (ed.), *Readings in the Philosophy of Psychology*, vol. 1 (Methuen, 1980); in D. Hofstadter and D. Dennett (eds.), *The Mind's I* (Harvester Press, 1981); and in Nagel's *Mortal Questions* (Cambridge University Press, 1979).

Ullan Place, 'Is consciousness a brain process?', *British Journal of Psychology*, vol. 47 (1956). Reprinted in C. Borst (ed.), *The Mind-Brain Identity Theory* (Macmillan, 1970); in V. Chappell (ed), *The Philosophy of Mind* (Prentice Hall, 1962); in A. Flew (ed.), *Body, Mind and Death* (Macmillan, 1964); and in H. Morick (ed), *Introduction to the Philosophy of Mind* (Harvester Press, 1979).

J. J. C. Smart, 'Sensations and brain processes', *Philosophical Review*, vol. 68 (1959). Reprinted in C. Borst (ed.), *The Mind-Brain Identity Theory* (Macmillan, 1970); in V. Chappell (ed), *The Philosophy of Mind* (Prentice Hall, 1962); and in D. Rosenthal (ed.), *Materialism and the Mind-Body Problem* (Prentice Hall, 1979).