6

Consciousness and the mechanical mind

The story so far

What should we make of the mechanical view of the mind?[1] In this book we have considered various ways in which the view has dealt with the phenomenon of mental representation, with our knowledge of the thoughts of others, and how (supplemented by further assumptions) it forms the philosophical basis of a computational view of thought. And, in the previous chapter, we looked at the attempts to explain mental representation in other terms, or 'reduce' it.

There are many questions unresolved: how adequate is the Theory Theory account of our understanding of others' thoughts? Do our minds have a connectionist or a classical 'architecture', or some combination of the two? Should a theory of mental representation attempt to reduce the contents of mental states to causal patterns of indication and the like, or is a non-reductive approach preferable? On some of these questions – e.g. connectionism vs. classicism – not enough is yet known for the sensible response to be other than a cautious open mind. On others – e.g. Theory Theory versus simulation – it seems to me that the debate has not yet been sharply enough formulated to know exactly what is at stake. It should be clear, though, that the absence of definite answers here should not give us reason to reject the mechanical view of the mind. For the essence of the mechanical view as I have characterised it is very hard to reject. It essentially involves commitment to the overwhelmingly plausible view that the mind is a causal mechanism which has its effects in behaviour. Everything else – computation, Theory Theory, reductive theories of content – is detail.

However, there are philosophers who do reject the view wholesale, and not because of the inadequacies of the details. They believe that the real problem with the mechanical view of the mind is that

it distorts – or even offers no account of – how our minds appear to us. It leaves out what is sometimes called the *phenomenology* of mind – where 'phenomenology' is the theory ('ology') of how things seem to us (the 'phenomena'). These critics object that the mechanical mind leaves out all the facts about how our minds strike us, what it feels like to have a point of view on the world. As far as the mechanical approach to the mind is concerned, they say, this side of having a mind might as well not exist. The mechanical approach treats the mind as 'a dead phenomenon, a blank agency imprinted with causally efficacious traces of recoverable encounters with bits of the environment'.[2] Or, to borrow a striking phrase of Francis Bacon's, the criticism is that the mechanical approach will 'buckle and bow the mind unto the nature of things'.[3]

In fact, something like this is a common element in some of the criticisms of the mechanical mind which we have encountered throughout this book. In Chapter 2, for instance, we saw that the Theory Theory was attacked by simulation theorists for its inadequate representation of what we do when we interpret others. By 'what we do when we interpret others', simulation theorists are talking about how interpretation *strikes* us. Interpretation does not *seem* to us like applying a theory – it's much more like an act of imaginative identification. (I do not mean to imply that simulation theorists are necessarily opposed to the whole mechanical picture; but they can be.) Yet why should anyone deny that interpretation sometimes seems to us like this? In particular, why should Theory Theorists deny it? And, if they shouldn't deny it, then what is the debate supposed to be about? The Theory Theory can reply that the issue is not how interpretation *seems* to us, but what makes interpretation *succeed*. The best explanation for the success of interpretation is to postulate tacit or implicit knowledge of a theory of interpretation. Calling this theory 'tacit' is partly to indicate that it is not phenomenologically available – that is, we can't necessarily tell by introspecting whether the theory is correct. But, according to the Theory Theory, this is irrelevant.

The same pattern of argument emerged when we looked at Dreyfus's critique of AI in Chapter 3. Dreyfus argued that thinking

cannot be a matter of manipulating representations according to rules. This is because thinking requires 'know-how', which cannot be reduced to representations or rules. But part of Dreyfus's argument for this is phenomenological: thinking does not seem to us like rule-governed symbol manipulation. It wouldn't be too much of a caricature to represent Dreyfus as saying: 'Just try it: think about some everyday task, like going to a restaurant, say – some task which requires basic cognitive abilities. Then try and figure out which rules you are following, and which "symbols" you are manipulating. You can't say what they are, except in the most open-ended and imprecise way'.

And, once again, the reply to this kind of objection on behalf of AI and the computational theory of cognition is that Dreyfus misses the point. For the point of the computational hypothesis is to explain the systematic nature of the causal transitions that constitute cognition. The computational processes that the theory postulates are not supposed to be accessible to introspection. So it cannot be an objection to the computational theory to say that we cannot introspect them.

In a number of debates, then, there seems to be a general kind of objection to mechanical hypotheses about the mind – that they leave out, ignore or cannot account for facts about how our minds seem to us, about the phenomenology of mind. In response, the mechanical view argues that how our minds seem to us is irrelevant to the mechanical hypothesis in question.[4]

It must be admitted that there is something unsatisfactory about this response. For the mechanical view cannot deny that there is such a phenomenon as how (our own and others') minds seem to us. And, what is more, many aspects of the idea of the mechanical mind are motivated by considering how the mind seems to us, in a very general sense of 'seems'. Consider, for example, the route I took in Chapter 2 from the interpretation of other minds to the hypothesis that thoughts are inner causal mechanisms, the springs of action. This is a fairly standard way of motivating the causal picture of thoughts, and its starting-points are common-sense observations about how we use conjectures about people's minds to explain their

behaviour. Another example is Fodor's appeal to the systematic nature of thought in order to motivate the Mentalese hypothesis. The examples that Fodor typically uses concern ordinary beliefs, as conceived by common sense: if someone believes that Anthony loves Cleopatra, then they must *ipso facto* have the conceptual resources to (at least) entertain the thought that Cleopatra loves Anthony. The starting points in many arguments for aspects of the mechanical mind are common-sense observations about how minds strike us. So it would be disingenuous for defenders of the mechanical mind to say that they have no interest at all in how minds seem to us.

The worry here is that, although it may start off in common-sense facts about how minds strike us, the mechanical view of the mind ends up saying things which seem to ignore how minds strike us, and thus depart from its starting point in common sense. What is the basis of this scepticism about the mechanical mind? Is it just that no defender of the view has yet come up with an account of the phenomenology of the mind? Or is there some deeper, more principled, objection to the mechanical mind which derives from phenomenology, which shows why the mechanical picture must be incorrect? In Chapter 5, we saw that many suppose that the *normativity* of the mental is one reason why a general reduction of mental representation must fail. The idea is that the facts that thought is true or false, correct or incorrect, that reasoning is sound or unsound, are all supposed to prevent an explanation of mental content in purely causal terms. But I argued that a conceptual reduction of mental content may not be essential to the mechanical picture of the mind. Representation may have to be considered a basic or fundamental concept in the theory of mind, without any further analysis. If this is true, then normativity is a basic or fundamental concept in the theory of mind too, because the idea of representation essentially carries with it the idea of correctness and incorrectness. But we saw no reason in this to deny that the underlying mechanisms of mental representation are causal in nature, and therefore no reason to deny the mechanical picture wholesale.

But there is another area in the investigation of the mind in which general arguments have been put forward that no causal

or mechanical picture of the mind can possibly give an adequate account of the phenomena of mind. This is the investigation into consciousness, postponed since Chapter 1. It is often said that consciousness is what presents the biggest obstacle to a scientific account of the mind. Our task in this chapter is to understand what this obstacle is supposed to be.

Consciousness, 'what it's like' and qualia

Consciousness is at once the most obvious feature of mental life and one of the hardest to define or characterise. In a way, of course, we don't need to define it. In everyday life, we have no difficulty employing the notion of consciousness – as when the doctor asks whether the patient has lost consciousness, or when we wonder whether a lobster is conscious in any way when it is thrown alive into a pan of boiling water. We may not have any infallible tests which will establish whether a creature is conscious or not; but it seems that we have no difficulty deciding what is at issue when trying to establish this.

Or at least, we have no difficulty deciding what is at issue as long as we don't try and reflect on what is going on. In considering the question, 'What is time?', Saint Augustine famously remarked that when no-one asks him, he knows well enough, but if someone were to ask him, then he does not know how to answer. The situation seems the same with 'What is consciousness?'. We are perfectly at home with the distinction between the conscious and the non-conscious when we apply it in ordinary life; but when we ask ourselves the question, '*What is consciousness?*', we are stuck for an answer. How should we proceed?

Well, what is the everyday distinction between the conscious and the non-conscious? We attribute consciousness to creatures, living organisms, and also to states of mind. People and animals are conscious; but so also are their sensations and (some of) their thoughts. The first use of the concept of consciousness has been called 'creature consciousness' and the second use 'state consciousness'.[5] Creature consciousness and state consciousness are obviously

interdependent: if a creature is conscious, that is when it is in conscious states of mind; and conscious states of mind are *ipso facto* the states of a conscious creature. There is no reason to suppose that we should define the one idea in terms of the other. But, nonetheless, it is perhaps easier to start our exploration of consciousness by considering what it is for a creature to be conscious. Thomas Nagel gave philosophers a vivid way of talking about the distinction between conscious and non-conscious creatures: a creature is conscious, he said, when there is something *it is like* to be that creature.[6] There is nothing it is like to be a bacterium, nothing it is like to be a piece of cheese – but something it is like to be a dog or a human being or (to use Nagel's famous example) a bat. This 'what it is like' idiom can be easily transferred to state consciousness too: there is something it is like to be tasting (to be in the state of tasting) vanilla ice-cream or to be smelling (to be in the state of smelling) burning rubber. That is, there is something it is like to be in these states of mind. But there is nothing it is like to be largely composed of water, or to have high blood pressure. These are not states of mind.

The phrase 'what it is like' is not supposed to be a definition of consciousness. But, as I have said already, we are not looking for a definition here. No-one lacking the concept of consciousness (if such a person were possible) would be able to grasp it by being told that there is something it is like to be conscious, or to be in conscious states. But we can say a couple of things about the meaning of this phrase which help to clarify its role in discussions of consciousness. First, the phrase is not intended in a *comparative* way. One might ask: what is Vegemite like? And the answer could be given: it's like Marmite. (For the uninitiated, Vegemite and Marmite are wonderful yeast-based condiments, the first from Australia, the second from the UK.) Here, asking what something is like is asking what things are *like it*; that is, what things resemble it. This is not the sense of 'what it's like' that Nagel intended when he said that there is something it is like to be a bat. Second, the phrase is not intended simply to mean *what it feels like*, if 'feels' has its normal meaning. For there are some states of mind where it makes sense to say that there is something it is like to be in these states, even

though this does not involve feeling in any ordinary sense. Consider the process of thinking through some problem, trying to understand some difficult task, in your head. There is, intuitively, something it is like to be thinking through this problem; but it need not 'feel' like anything. There need be no special feelings or sensations involved. So, although there is something it is like to feel a sensation, not all cases where there is something it is like are cases of feelings.

'What it is like', then, does not mean *what it resembles* and it does not (just) mean *what it feels like*. What it is trying to express is how things seem to us when we are conscious, or in conscious states, what I called in the previous section the *appearance* or the *phenomena* of mind. This is supposed to be different from merely being the kind of creature which has a mind: What it *is* to be a bat is one thing; what it is *like* to be a bat is another. Now, the term 'phenomenal consciousness' is sometimes used for this idea of how things seem to a conscious creature; and the term is etymologically apt, given that the English word 'phenomenon' is derived from the Greek word for *appearance*. A creature is phenomenally conscious when there is something it is like to be that creature; a state of mind is phenomenally conscious when there is something it is like to be in that state. The special way a state of mind is, what constitutes what it is like to be in that state, is likewise called the *phenomenal character* of the state.

Sometimes phenomenal consciousness is described in terms of *qualia* (we first encountered qualia in Chapter 1, 'Brentano's thesis'). Qualia (plural: the singular is *quale*) are supposed to be the non-representational, non-intentional, yet phenomenally conscious properties of states of mind.[7] Believers in qualia say that the particular character of the aroma of smelling coffee cannot just be captured in terms of the way the smell represents coffee; this would fail to capture the way it *feels* to smell coffee. Even when you have described all the ways your experience of the smell of coffee represents coffee, you will have left something out: that is the qualia of the experience of smelling coffee, the *intrinsic properties* of the experience, which are independent of the representation of coffee. Someone who believes in qualia denies Brentano's thesis that all

mental phenomena are intentional: certain conscious properties of states of mind are not intentional at all. And these are supposed to be the properties which are so hard to make sense of from a naturalistic point of view. Hence the problem of consciousness is often called the 'problem of qualia'.[8]

But, though it is not controversial that there is such a thing as phenomenal consciousness, it *is* controversial that there are qualia. Some philosophers deny that there are any qualia, and by this they do not mean that there is no phenomenal consciousness.[9] What they mean is that there is nothing to phenomenal consciousness over and above the representational properties of states of mind. In the case of visual perception, for example, these philosophers – known as *intentionalists* or *representationalists* – say that when I perceive something blue I am not aware of some *intrinsic* property of my state of mind, in addition to the blueness which I perceive. I look at a blue wall, and all I am aware of is the wall and its blueness. I am not, in addition, aware of some intrinsic properties of my state of mind.[10] And this view says similar things about sensation. The believer in qualia says that, in such a case, one is also aware of what Ned Block has called 'mental paint': the intrinsic properties of one's state of mind.

Things can become confusing here because other philosophers use the word 'qualia' simply as a synonym for 'phenomenal character' – so that to have phenomenal consciousness is, as a matter of definition, to have qualia. This is very unhelpful because it makes it impossible to understand what philosophers such as Tye and Dennett could possibly mean when they deny that there are qualia. To make a first attempt at clarifying matters here, we must distinguish two ways of using the term 'qualia': (i) to have qualia is simply to have experience with a phenomenal character; or (ii) qualia are non-intentional (non-representational) qualities of experience.

The debate about consciousness involves, it seems, a large amount of terminological confusion. We need to make a broad distinction between phenomenal consciousness – the thing to be explained – and those properties that are appealed to in order to explain phenomenal consciousness. Unless we do this we will not understand

what it is that philosophers are doing when they deny the existence of qualia. Superficially, it might look as if they are rejecting the phenomena of consciousness, whereas what they are really rejecting is a certain way of explaining phenomenal consciousness: in terms of qualia, non-intentional, non-representational properties of mental states.

These clarifications made, we must finally turn to an overdue topic, the mind–body problem.

Consciousness and physicalism

In Chapter 2 ('The mind–body problem') I said that the mind–body problem can be expressed in terms of the puzzlement which we feel when trying to understand how a mere piece of matter like the brain can be the source of something like consciousness. On the one hand, we feel that our consciousness must just be based on matter; but, on the other hand, we find it impossible to understand how this can be so. This is certainly what makes many people think that consciousness is mysterious; but, by itself, it is not a precise enough thought to give rise to a philosophical problem. Suppose someone were to look at a plant, and having found out about the processes of photosynthesis and cellular growth in plants, still found it incredible that plants could grow only with the help of sun, water and soil. Tough. No interesting philosophical consequences should be drawn from this person's inability to understand the scientific facts. Of course, life and reproduction can look like remarkable and mysterious phenomena; but the proper response to this is simply to accept that certain phenomena in nature are remarkable and maybe even mysterious. But that doesn't mean that they cannot be explained by science. The ability of creatures to reproduce themselves is now fairly well understood by scientists; it may be remarkable and mysterious for all that.

To approach the issue in another way, consider the argument that physicalist or materialist views typically give for their view that mental states (both thoughts and conscious states) are identical with states of the brain. In rough outline, they argue, first, that conscious

and other mental states have effects in the physical world (perhaps using the kinds of argument which I used in Chapter 2, 'The causal picture of thoughts', p. 54); and, second, that every physical happening is the result of purely physical causes, according to physical law (this is sometimes called 'the causal closure of the physical').[11] I cannot go into the reasons for this second assumption in any detail here. Let's just say that physicalists believe that this is the consequence of what we have learned from science: science succeeds in its explanatory endeavours by looking for the underlying mechanisms for things which happen. And looking for the underlying mechanisms ends up uncovering physical mechanisms – the sorts of mechanisms discovered in physics, the science of spacetime, matter and energy. As David Lewis puts it:

> [T]here is some unified body of scientific theories of the sort we now accept, which together provide a true and exhaustive account of all physical phenomena. They are unified in that they are cumulative: the theory governing any physical phenomenon is explained by theories governing phenomena out of which that phenomenon is composed and by the way it is composed out of them. The same is true of the latter phenomena, and so on down to fundamental particles or fields governed by a few simple laws, more or less as conceived in present-day theoretical physics.[12]

It is this kind of thing which grounds physicalists' confidence in the idea that, ultimately, all physical effects are the result of physical causes. They then conclude that, if mental causes really do have effects in the physical world, then they must themselves be physical. For. if mental causes weren't physical, then there would be physical effects which are brought about by non-physical causes, which contradicts the second assumption.

This is a quite general argument for identifying mental states with physical states (for example, states of the brain). Call it the 'causal argument for physicalism'. Although it rests on a scientific or empirical assumption about the causal structure of the physical world, the causal argument for physicalism does not rely on scientists actually having discovered the basis in the brain (what

they tend to call the 'neural correlate'[13]) of any particular mental state. Although most physicalists think that such neural correlates will eventually be found, they are not *presupposing* that they will be found; all they are presupposing in this argument is the causal nature of mental states and the causal closure of the physical world. It follows that one could object to the conclusion of the argument either by objecting to the causal nature of mental states, or by objecting to the causal closure of the physical world, or by saying that there is some confusion or fallacy in moving from these two assumptions to the conclusion that mental states are states of the brain.

But notice that it is not a serious objection to this conclusion just to say: 'but mental states do not *seem* to be states of the brain!'. This is, it must be admitted, a very natural thought. For it is true that when one introspects one's states of mind – in the case of trying to figure out what one is thinking, for example – it does not *seem* as if we are obtaining some sort of direct access to the neurons and synapses of our brains. But, if the argument above is right, then this evidence from introspection is irrelevant. For if it *is* true that mental states are states of the brain, then it will be true that, as a matter of fact, being a certain brain state will seem to you to be a certain way, although it might not seem to be a brain state. But that's OK; it can seem to you that George Orwell wrote *1984* without its seeming to you that Eric Blair did, even though, as a matter of fact, Eric Arthur Blair did write *1984*. (Logicians will say that 'it seems to me that ...' is an *intensional context*: see Chapter 1, 'Intentionality', p. 30.) The conclusion of the causal argument for physicalism is that mental states are brain states. To object to this by saying, 'but surely mental states can't be brain states, because they don't seem to be!' is not to raise a genuine objection: it is just to reject the conclusion of the argument. It is as if someone said, in response to the claim that matter is energy, 'matter cannot be energy because it does not seem like energy!'. In general, when someone asserts some proposition, P, it is not a real objection to say, 'but P does not seem to be true; therefore it is not true!'. And the point is not that one might not be *correct* in denying P. The point is rather that there is a distinction between raising an objection to a thesis and denying the thesis.

So mental states might be brain states, even if they do not seem to be. We can illustrate this in another way, by using a famous story about Wittgenstein. 'Why did people used to think that the sun went around the earth?' Wittgenstein once asked. When one of his students replied 'Because it looks as if the sun goes around the earth', he answered, 'And how would it look if the earth went around the sun?'. The answer, of course, is: exactly the same. So we can make a parallel point in the case of mind and brain: why do some people think that mental states are not brain states? Answer: because mental states do not seem like brain states. Response: but how would they seem if they were brain states? And the answer to this, of course, is: exactly the same. Therefore, there is no simple inference from the fact that being in a mental state makes things seem a certain way to any conclusion about whether mental states have a physical nature or not.

No *simple* inference; but maybe there is a more complicated one concealed inside this (admittedly very natural) objection. Some philosophers think so; and they think that it is *consciousness* which really causes the difficulty for physicalism (and, as we shall see, for the mechanical mind too). There are various versions of this problem of consciousness for physicalism. Here I will try and extract the essence of the problem; the Further reading section (pp. 231–232) will indicate ways in which the reader can explore it further.

The essence of the problem of consciousness derives from the apparent fact that any physicalist description of conscious states seems to be, in Nagel's words, 'logically compatible with the absence of consciousness'. The point can be made by comparison with other cases of scientific identifications – identifications of everyday phenomena with entities described in scientific language. Consider, for example, the identification of water with H_2O. Chemistry has discovered that the stuff that we call 'water' is made up of molecules which are themselves made up of atoms of hydrogen and oxygen. There is nothing more to being water than being made up of H_2O molecules; this is why we say that water *is* (i.e. *is identical with*) H_2O. Given this, then, it is not logically possible for H_2O to exist and water not to exist; after all, they are the same thing! Asking whether

there could be water without H_2O is like asking whether there could be George Orwell without Eric Arthur Blair. Of course not; they are the same thing.

If a conscious mental state – for example, a headache – were really identical with a brain state (call it 'B' for simplicity), then it would in a similar way be impossible for B to exist and for the headache not to exist. For, after all, they are supposed to be the same thing. But this case does seem to be different from the case of water and H_2O. For whereas the existence of water without H_2O seems absolutely impossible, the existence of B without the headache does seem to be possible. Why? The short answer is: because we can coherently conceive or imagine B existing without the headache existing. We can conceive, it seems, a creature who is in all the same brain states as I am in when I have a headache but who in fact does not have a headache. Imaginary creatures like this are known in the philosophical literature as 'zombies': a zombie is a physical replica of a conscious creature who is not actually conscious.[14] The basic idea behind the zombie thought-experiment is that, although it does not seem possible to have H_2O without water, it does seem possible (because of the possibility of zombies) to have a brain state without a conscious state; so consciousness cannot be identical with or constituted by any brain states.

This seems like a very fast way to refute physicalism! However, although it is very controversial, the argument (when spelled out clearly) does not involve any obvious fallacy. So let's spell it out more slowly and clearly. The first premise is:

1 If zombies are possible, then physicalism is false.

As we saw in Chapter 1, physicalism has been defined in many ways. But here we will just take it to be the view that is the conclusion of the causal argument above: mental states (including conscious and unconscious states) are identical with states of the brain. The argument against physicalism is not substantially changed, however, if we say that, instead of being identical with states of the brain, mental states are exhaustively *constituted* by

states of the brain. Identity and constitution are different relations, as identity is *symmetrical* where constitution is not (see Chapter 1: 'Pictures and resemblance', p. 13, for this feature of relations). If Orwell *is identical with* Blair, then Blair *is identical with* Orwell. But if a parliament *is constituted by* its members, then it does not follow that the members *are constituted by* parliament. Now, one could say that states of consciousness are constituted by states of the brain, or one could say that they are identical with states of the brain. Either way, the first premise does seem to be true. For both ideas are ways of expressing the idea that conscious states are *nothing over and above* states of the brain. Putting it metaphorically, the basic idea is that, according to physicalism, all God needs to do to create my conscious states is to create my physical brain. God does not need to add anything else. So, if it could be shown that creating my brain is not enough to create my states of consciousness, then physicalism would be false. Showing that zombies are possible is a way of showing that creating my brain is not enough to create my states of consciousness. This is why premise 1 is true.

The next premise is:

2 Zombies are conceivable (or imaginable).

What this means is that we can coherently imagine a physical replica of a conscious being (e.g. me) without any consciousness at all. This zombie-me would have all the same physical states as me, the same external appearance, and the same brain and so on. But he would not be conscious: he would have no sensations, no perceptions, no thoughts, no imagination, nothing. Perhaps we can allow him to have all sorts of unconscious mental states (the sort described in Chapter 1, 'Thought and consciousness', p. 26). But what he has nothing of is consciousness of any kind. Obviously, when we are imagining the zombie, we are imagining it from the 'outside'; we cannot imagine it from the 'inside', from the zombie's own point of view. For there is, of course, no such thing as the zombie's point of view.

Let's just be clear about what premise 2 says. If someone asserts

premise 2, they are not saying that there *really are any zombies*, or that *for all I know, you might all be zombies*, or that they are possible in any *realistic or scientific* sense. Not at all. One can deny outright that there are any zombies, deny that I have any doubts about whether you are conscious, and deny that there could be, consistent with the laws of nature as we know them, any such things – and one can still hold premise 3. Premise 3 asserts the mere, bare possibility of physical replicas who are not conscious.

There is no obvious contradiction in stating the zombie hypothesis. But maybe there is an unobvious one, something hidden in the assumptions we are making, which shows why premise 2 is really false. Perhaps we are merely *thinking* that we are imagining the zombie, but we aren't really coherently imagining anything. It can happen that someone tries to imagine something, and seems to imagine it, but does not really succeed in imagining precisely *that thing* because it is not really possible. I might, for example, try and imagine being my brother. I think I can imagine this, living where he is living, doing what he is doing. But of course I cannot literally *be* my brother: no-one can literally *be identical* with someone else. This is impossible. So maybe I am failing to imagine literally being my brother, and really imagining something else. Maybe what I am really imagining is me, myself, living a life rather like my brother's life. We can say a similar thing about the parallel case of water and H_2O: someone might think that they can imagine water not being H_2O, but having some other chemical structure. But, arguably, they are not really imagining *this*, but rather imagining something that looks just like water, but isn't water (as water is, by hypothesis, H_2O).[15] So someone can fail to imagine something because it is impossible: premise 2 might be false.

There is, however, another way of criticising the argument: we could agree that my being my brother is impossible; but all this shows is that one can imagine impossible things. In other words, we could accept the first two premises in this argument, but reject the move from there to the next premise:

3 Zombies are possible.

Obviously, 3 and premise 1 imply the conclusion:

4 Physicalism is false.

So anyone who wants to defend physicalism should concentrate on the key point in the argument, the move from premise 2 to premise 3. How is this move supposed to go? Premise 2 is supposed to provide the reason to believe in premise 3. The argument says that we should believe in premise 3 because of the truth of premise 2. Notice that it is one thing to say that if X is conceivable then X is possible, and quite another to say that being conceivable is the *same thing* as being possible. This is implausible. Some things may be imaginable without being really possible (e.g. someone might imagine a counterexample to a law of logic), and some things are possible without being imaginable (for example, for myself, I find it impossible to imagine or visualize curved spacetime). Imaginability and possibility are not the same thing. But they are related, according to this argument: imaginability is the best evidence there is for something's being possible. Rather as perception stands to what is real, so imagination stands to what is possible. Perceiving something is good evidence that it is real; imagining something is good evidence that something is possible. But the real is not just the perceivable, just as the possible is not just the imaginable.

The physicalist will respond to this that while it may be true in general that the imagination is a good guide to possibility, it is not infallible, and it can lead us astray (remember the Churchlands' example of the luminous room in Chapter 3, 'The Chinese Room', p. 123). And they would then argue that the debate about consciousness and zombies is an area where it does lead us astray. We imagine something, and we think it possible; but we are misled. Given the independent reasons provided for the truth of physicalism (the causal argument above), we know it cannot be possible. So what we can imagine is, strictly speaking, irrelevant to the truth of physicalism. That's what the physicalist should say.

To take stock: there are two ways a physicalist can respond to the zombie argument. The first is to deny premise 2 and show that

zombies are not coherently conceivable. The second is to accept 2 and reject the move from 2 to 3. So, for the physicalist, either zombies are inconceivable and impossible, or they are conceivable but impossible. It seems to me that the second line of attack is less plausible: for if physicalists agree that, in some cases, imaginability is a good guide to possibility, then what is wrong with this particular case? Physicalists would be better off taking the first move, and attempt to deny that zombies are really, genuinely conceivable. They have to find some hidden confusion or incoherence in the zombie story. My own view is that there is no such incoherence; but the issues here are very complicated.

The limits of scientific knowledge

But suppose that the physicalist can show that there is a hidden confusion in the zombie story – maybe zombies are kind of conceivable, but not really possible. So the link between the brain and consciousness is necessary, appearances to the contrary. Still physicalism is not home and dry. For there are arguments, related to the zombie argument, which aim to show that, even if this were the case, physicalism would still have an epistemological shortcoming: there would nonetheless be things which physicalism could not explain. Even if physicalism were metaphysically correct – correct in the general claims it makes about the world – its account of our knowledge of the world will be necessarily incomplete.

The easiest way to see this is to outline briefly a famous argument, expressed in the most rigorous form in recent years by Frank Jackson: he called it 'the knowledge argument'. [16] Let's put the argument this way. First, imagine that Louis is a brilliant scientist who is an absolute expert on the physics, physiology and psychology of taste, and on all the scientific facts about the making of wine, but has never actually tasted wine. Then one day Louis tastes some wine for the first time. 'Amazing!' he says, 'so this is what Chateau Latour tastes like! Now I know.'

This little story can then provide the basis of an argument with two premises:

1 Before he tasted wine, Louis knew all the physical, physiological, psychological and enological facts about wine and tasting wine.
2 After he tasted wine, he learned something new: what wine tastes like.

Conclusion: Therefore, not everything that there is to know about tasting wine is something physical. There must therefore be non-physical things to learn about wine: *viz*. what it tastes like.

The argument is intriguing. For, if we accept the coherence of the imaginary story of Louis, then the premises seem to be very plausible. But the conclusion does seem to follow, fairly straightforwardly, from the premises. For if Louis did learn something new then there must be something that he learned. You can't learn without learning something. And, because he already knew all the physical things that there are to know about wine and wine-tasting, the new thing he learns cannot be something physical. But if this is true then it must be that not everything we can know falls within the domain of physics. And not just physics: any science whatsoever that one could learn without having the experiences described by that science. Jackson concluded that physicalism is false: not everything is physical. But is this right?

The argument is very controversial, and has inspired many critical responses. Some people don't like thought-experiments like the story of Louis.[17] But it's really hard to see what could possibly be wrong with the idea that, when someone drinks wine for the first time, they come to learn something new: they learn what it tastes like. So, if we were going to find something wrong with the story itself, it would have to be with the idea that someone could know *all* the physical facts about wine and wine tasting. True enough, it is hard to imagine what it would be to learn all these facts. As Dennett says, you don't imagine someone having all the money in the world by imagining them being very rich.[18] Well, yes; but if you really do want to imagine someone having all the money in the world, you surely wouldn't go far wrong if you started off imagining

them being very very rich and then more so, without ever having to imagine them having more of anything of a *different kind*, just more of the same: money. And likewise with scientific knowledge: we don't have to imagine Louis having anything of a very *different kind* from the kind of scientific knowledge that people have today: just more of the same.

The standard physicalist response to the argument is rather that it doesn't show that there are any non-physical *entities* in the world. It just shows that there is non-physical *knowledge* of those entities. The *objects* of Louis's knowledge, the physicalist argues, are all perfectly ordinary physical things: the wine is made up of alcohol, acid, sugar and other ordinary physical constituents. And we have not been shown anything which shows that the change in Louis's subjective state is anything more than a change in the neurochemistry of his brain. Nothing in the argument, the physicalist claims, shows that there are any non-physical objects or properties, in Louis's brain or outside it. But they do concede that there is a change in Louis's state of knowledge: he knows something he did not know before. However, all this means is that states of knowledge are more numerous than the entities of which they are knowledge. (Just as we can know the same man as Orwell and come to know something new when we learn he is Blair.)

But this is not such a happy resting place for physicalists as they might think. For what this response concedes is that there are, in principle, limits to the kind of thing which physical science can tell us. Science can tell us about the chemical constitution of wine; but it can't tell us what wine tastes like. Physicalists might say that this is not a big deal; but, if they do say this, they have to give up the idea that physics (or science in general) might be able to state every *truth* about the world, independently of the experiences and perspectives of conscious, thinking beings. For there are truths about what wine tastes like, and these are the kind of truths you can only learn having tasted wine. These are truths which Louis would not have learned before tasting wine, I believe, no matter how much science he knew. So there are limits to what science can teach us – though this is a conclusion which will only be surprising or

disturbing to those who thought that science could tell us everything in the first place.

So let's return finally to the mind–body problem. Contrary to what we might have initially thought, the problem can now be clearly and precisely formulated. The form of the problem is that of a dilemma. The first horn of the dilemma concerns mental causation: if the mind is not a physical thing, then how can we make sense of its causal interactions in the physical world? The causal argument for physicalism says that we must therefore conclude that the mind is identical with a physical thing. But the second horn of the dilemma is that, if the mind is a physical thing, how can we explain consciousness? Expressed in terms of the knowledge argument: how can we explain what it *feels* like to taste something, even if tasting something is a purely physical phenomenon? Causation drives towards physicalism, but consciousness drives us away from it.

Conclusion: what do the problems of consciousness tell us about the mechanical mind?

What does the mind–body problem have to do with the mechanical mind? The mechanical view of the mind is a causal view of mind; but it is not necessarily physicalist. So an attack on physicalism is not necessarily an attack on the mechanical mind. The heart of the mechanical view of the mind is the idea that the mind is a causal mechanism which has its effects in behaviour. Mental representation undoubtedly has causal powers, as we saw in Chapter 2, so this relates the mechanical mind directly to the mind–body problem. We have found no good reason, in our investigations in this book, to undermine this view of representation as causally potent. But the mechanical view still has to engage with the causal argument for physicalism outlined in this chapter; and, if a physicalist solution is recommended, the view has to say something about the arguments from consciousness which form the other half of the dilemma which is the mind–body problem. Given the close inter-relations between thought and consciousness, the question of consciousness cannot be

ignored by a defender of the mechanical mind. (Fodor, characteristically, disagrees: 'I try never to think about consciousness. Or even to write about it.'[19]) The positive conclusion is that we have unearthed no powerful argument against the view that the mind is a causal mechanism which has its effects in behaviour.

Nonetheless, our investigations into the mechanical mind have also yielded one broad and negative conclusion: there seems to be a limit to the ways in which we can give *reductive* explanations of the distinctive features of the mind. We found in Chapter 3 that, although there are interesting connections between the ideas of computation and mental representation, there is no good reason to suppose that something could think simply by being a computer: reasoning is not just reckoning. In Chapter 4, we examined the Mentalese hypothesis as an account of the underlying mechanisms of thought; but this hypothesis does not reductively explain mental representation, but takes it for granted. The attempts to explain representation in non-mental terms examined in Chapter 5 foundered on some fundamental problems about misrepresentation and complexity. And, finally, in the present chapter, we have seen that, even if the attacks on physicalism from the 'conceivability' arguments are unsuccessful, they have variants which show that there are fundamental limits to our scientific knowledge of the world. Perhaps the proper lesson should be that we should try and be content with an understanding of mental concepts – representation, intentionality, thought and consciousness – which deals with them in their own terms, and does not try and give reductive accounts of them in terms of other sciences. And perhaps this is a conclusion which, in some sense, we already knew. Science, Einstein is supposed to have remarked, cannot give us the taste of chicken soup. But – when you think about it – wouldn't it be weird if it did?

Further reading

An excellent collection of essays on the philosophy of consciousness is *The Nature of Consciousness* edited by Ned Block, Owen Flanagan and Güven Güzeldere (Cambridge, Mass.: MIT Press 1997). This contains Thomas Nagel's

classic paper, 'What is it like to be a bat?', Colin McGinn's 'Can we solve the mind–body problem?', Jackson's 'Epiphenomenal qualia', Block's 'On a confusion about a function of consciousness' and many others. See also *Conscious Experience* edited by Thomas Metzinger (Paderborn: Schöningh 1995). Much of the agenda in recent philosophy of consciousness has been set by David Chalmers's ambitious and rigorous *The Conscious Mind* (New York, NY and Oxford: Oxford University Press 1996). Joseph Levine's *Purple Haze* (New York, NY and Oxford: Oxford University Press 2001) gives a very clear, though ultimately pessimistic, account of the problem of consciousness for materialism, in terms of what Levine has christened the 'explanatory gap'. David Papineau's *Thinking About Consciousness* (Oxford: Oxford University Press 2002) is a very good defence of the view that the problems for physicalism lie in our concepts rather than in the substance of the world. On the debate over intentionality and qualia, Michael Tye's *Ten Problems of Consciousness* (Cambridge, Mass.: MIT Press 1995) is a good place to start. Daniel Dennett's *Consciousness Explained* (London: Allen Lane 1991) is a philosophical and literary tour de force, the culmination of Dennett's thinking on consciousness; controversial and hugely readable, no philosopher of consciousness can afford to ignore it. Gregory McCulloch's *The Life of the Mind* (London and New York: Routledge 2003) offers an unorthodox non-reductive perspective on these issues.

THE MECHANICAL MIND

A philosophical introduction to
minds, machines and mental representation

SECOND EDITION

TIM CRANE

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 1995
by Penguin Books
Second edition published 2003
by Routledge
11 New Fetter Lane, London EC4P 4EE

Simultaneously published in the USA and Canada
by Routledge
29 West 35th Street, New York, NY 10001

Routledge is an imprint of the Taylor & Francis Group

This edition published in the Taylor & Francis e-Library, 2003.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
A catalog record for this book has been requested

ISBN 0-203-42631-2 Master e-book ISBN

ISBN 0-203-43982-1 (Adobe eReader Format)
ISBN 0-415-29030-9 (hbk)
ISBN 0-415-29031-7 (pbk)