contingent identity as if it were a special, contingent kind of identity that held between, for instance, pain and *C* fibres firing, and curare and the most insidious poison. But, of course, in each case there is just *one* thing and it is necessarily self-identical.

Some have objected to the anti-essentialist nature of the mind–brain identity theory. They have insisted that mental states are the kind of mental states that they are essentially. Pain could not have been anything other than pain; belief that snow is white could not have been anything other than belief that snow is white. However, provided that we hold to the position that mental states play distinctive causal roles, causal roles that figure centrally in determining the kind of mental state that they are, essentialism about mental states is hard to sustain. What a state does is *not* an essential property of it. Physicalists sympathetic to functionalism have a choice to make here. In the terms we introduced above, only if they hold that mental states are realizer states, not role states, can they give them their intuitively plausible causal roles; but then mental states are not the mental states they are essentially.

Annotated Reading

The classic article-length presentation of the identity theory is J. J. C. Smart, 'Sensations and Brain Processes'. The classic book-length presentation is D. M. Armstrong, *A Materialist Theory of the Mind*. Although both these works are directed to fellow professionals, they are written in a very direct and clear way that makes them highly accessible. A paper which is very clear about the irrelevance of Occam's Razor and is explicitly a defence of the type–type version of the theory is David Lewis, 'An Argument for the Identity Theory'. The advantages of the type–type version over the token–token version are spelt out in a little detail in Frank Jackson, Robert Pargetter and Elizabeth Prior, 'Functionalism and Type–Type Identity Theories'. More elementary presentations of the identity theory can be found in most philosophy of mind texts, though typically, as we have said, the theory is presented before functionalism, and especially in American texts, as a view superseded by functionalism. Recent discussions of whether essentialist considerations make trouble for the identity theory date from lecture 3 in Saul Kripke, *Naming and Necessity*.

7

FOUR CHALLENGES TO FUNCTIONALISM

If functionalism is true, anything that is functionally like us in the relevant respects is psychologically like us. This chapter is concerned, first, with three well-known examples of things that are, in one way or another, and to one extent or another, functionally like us and yet which intuitively are very unlike us psychologically. We will consider in turn the challenge posed to functionalism by the China Brain, the Chinese room, and Blockhead. In each example we consider (a) whether the example really is, on reflection, of a creature that is very unlike us psychologically, and (b) to the extent that it is, we ask whether the example shows that functionalism is false, or does it instead teach us something important about *which* functional roles are crucial? We will suggest that to the extent that the examples are indeed of entities that are psychologically unlike us, they are examples of entities that do not have the right functional similarities to us, and so tell us not that functionalism is false but rather about the functional roles that functionalists need to include in their story about psychological nature. We will need to advert at various points to the fact that functionalism comes in many flavours in our discussion of the three examples. Fortunately, however, the examples raise many rather general issues about functionalism, and this means that we will often be able to think in terms of a fairly undifferentiated version that captures general features common to most versions.

In the final section we will discuss a general objection to any kind of functionalism or indeed physicalism from the alleged possibility of 'zombies'.

The China Brain

The China Brain is a putative counter-example to functionalism due originally to Ned Block. Here is a slightly updated version of the example. Imagine that artificial intelligence has advanced to the point where a program can be written which will allow an android with a 'brain' consisting of a computer running the program to behave actually and counterfactually much as a normal human does. It does not matter for the example how this programming is done; to avoid confusion about the nature of the program (which we will discuss in a later example), let us suppose that the program mimics the operation of a human brain at a neuron by neuron level. Neurons are essentially 'input–output' devices made from organic matter, the overall input–output characteristics of the brain being determined by how the primitive neuronal devices are assembled. Hence, this supposition amounts to having the program reflect precisely the input–output nature of each neuron and how they are connected one to another.

The next step in the process of constructing the example is to note that it won't matter, or anyway can hardly matter from a functionalist perspective, if the computer running this program is in fact outside the android's body, connected by a two-way radio link to it. The final step gives us the China Brain. Suppose that instead of the program being run on an external computer made of silicon chips, the entire population of China is enlisted to run the simulation. As the program mimics the way the brain operates at the neuronal level, this can be done by assigning each Chinese citizen the job of just one neuron. They have, let's suppose, the kind of phones that tell you what number has called you. When certain numbers, or combinations of numbers, ring in, they have to dial specified other numbers. Each citizen is given a precise set of instructions about what to do that ensures that what each does exactly models what their assigned neuron does, and the inputs to and outputs from their phones are connected up so as to run the program. Also, the initial inputs to the China brain come from the environment in much the same way as the inputs to us do, and the final outputs go to the limbs and head of the android via the radio link in such a way that its actual and counterfactual behaviour is much as ours is. Thus, the android will behave in the various situations that confront it very much as we do, despite the fact that the processing of the environmental inputs into final behavioural outputs goes via a highly organized set of Chinese citizens rather than a brain.

This is certainly not a realistic fantasy. The population of China is not large enough; the whole process could never take place fast enough; the citizens would get bored and careless; and anyway the program used to construct the example does not exist and never will (working at the neuronal level is ridiculously fine-grained). All the same it does seem clearly intelligible, and if it is intelligible, it is fair to ask for an answer to the question whether the system consisting of the robot plus the population of China in the imagined case has mental states like ours. Many have a strong

intuition that it does not. If they are right, functionalism of just about any variety must be false. For the system is functionally very like us. Not only is it like us in all the functional roles seen as crucial by the common-sense functionalist, it is like us in just about every functional respect. Functionally, it *is* us; the difference lies in the dramatic difference in how the functional roles are realized, and that difference counts for nothing as far as mental nature is concerned according to functionalists.

Denying the intuition

We think, however, that the functionalist can reasonably deny the intuition. The source of the intuition that the system consisting of robot plus China brain lacks mental states like ours seems to be the fact that it would be so *very* much bigger than we are. We cannot imagine 'seeing' it as a cohesive parcel of matter. We cannot see, that is to say, the forest for the trees. A highly intelligent microbe-sized being moving through our flesh and blood brains might have the same problem. It would see a whole mass of widely spaced entities interacting with each other in a way that made no sense to it, that formed no intelligible overall pattern from its perspective. The philosophers among these tiny beings might maintain with some vigour that there could be no intelligence here. All that is happening is an inscrutable sending back and forth of simple signals. They would be wrong. We think that the functionalist can fairly say that those who deny mentality in the China brain example are making the same mistake.

Consciousness

Before we leave the China brain example, we should note two important points about its role in the literature. First, it is sometimes directed simply to the question of whether functionalism can account for consciousness. In this manifestation it is granted that the China brain has beliefs and desires (after all, the robot will move in various ways in response to the environment and thereby make changes to it of just the kind we associate with purposive, informed behaviour), but it is insisted that it is absurd to hold that it *feels* anything. We discuss the difficult question of feeling and consciousness in the next chapter. Our concern in this chapter will be restricted to challenges for functionalism about mental states like belief and desire, and mental traits like being intelligent.

Connection to the environment

Secondly, sometimes the example is given in a version that omits the robot. But then the population of China is emulating, in some purely abstract way, the program in someone's brain with no obvious right way to connect the overall inputs and outputs with the environment. The case becomes essentially the same as the one we discussed when we considered the charge of excessive liberalism against certain machine versions of functionalism in chapter 5. There we argued that merely crunching numbers – or more generally inputs and outputs that have no natural or obvious connection to the environment that our mental states are about – is *mere* number (or whatever) crunching. If this is right, then the China brain example in the version that omits the appropriate robotic connection to the environment is an example of something that lacks a mental life, or anyway a mental life at all like ours. But in this version it is not an objection to functionalisms that include, in one form or another, the right sort of connection to the environment. These functionalisms might be called 'arm's length' functionalisms. Common-sense functionalism is an 'arm's length' doctrine in this sense.

The Chinese Room

John Searle's Chinese room is one of the most famous examples in the philosophy of mind. We will present a variation on the original example. We suppose that someone called Tex, who understands English but not Chinese, is locked in a room that has an in-chute, an out-chute and a book full of instructions in English concerning the manipulation of Chinese characters. Stories in Chinese accompanied by questions in Chinese about the stories come in through the in-chute. Tex follows the instructions in the book as applied to the stories and the questions, which in due course tell him which sentences in Chinese to copy onto pieces of paper and place in the out-chute. Tex does not understand the stories, the questions or the sentences he puts in the out-chute. As far as he is concerned he is simply operating with squiggles. He is mechanically following some rules in English for manipulating symbols which are in various ways derived from the squiggles, and which conclude with his writing down some more squiggles on the paper that goes into the out-chute. We can make the example more up to date (and facilitate later variations on it that we will discuss) by supposing the stories and questions in Chinese are typed into a computer outside the room and appear on its screen. Tex has a monitor in his room on which what is typed is also displayed in Chinese. He types in answers to the questions at a separate keyboard in the room, following the book's instructions religiously. His answers appear both on his monitor, and on the monitor screen outside the room. We will conduct the discussion in terms of this version of the example.

Searle in effect points out that the book might well be such that Tex will consistently deliver Chinese sentences that, to someone who understands Chinese, count as sensible, intelligent answers to the questions in Chinese about the stories in Chinese that Tex receives. What appears on the screen are good answers in Chinese to questions in Chinese about stories in Chinese. Nevertheless, it would be quite wrong to infer that Tex understands Chinese. All he is doing is manipulating symbols according to formal rules without any understanding of what the various symbols stand for or mean.

It is clear that Tex does not understand Chinese, for he does not himself have the ability to answer the questions. It is Tex together with the book that has the ability. So the issue that needs to be addressed is whether the *system* consisting of Tex plus the book understands Chinese. What abilities are distinctively associated with understanding a language? It is plausible that being able to answer comprehension questions about a range of stories is part of what is required, but is very much less than all that is required. One thing we need to add is the ability to extemporize, embellish, and generally display the inventiveness and flexibility of a natural language speaker. If we *always* get back the same answer – accurate and intelligent though it may seem – in response to a story together with a question, we might well start to think that we are interacting with an automaton rather than a thinker and understander.

The example embellished

We can, though, embellish the original example by supposing that Tex plus room has all these capacities. We can suppose that the book Tex is following does not always deliver the same answer in Chinese to a given question in Chinese. The book takes account of whether or not a question has been asked before. It contains instructions in English concerning what to do when a given sequence of squiggles (as Tex thinks of them) appears on the computer screen that takes into account whether and how often that sequence of Chinese characters has appeared before. Obviously, by making Tex's book of instructions sufficiently complex (it will have to include instructions on how to modify the book itself in response to input, even if only by leaving different pages open), we can ensure that the answers Tex generates on the screen are exactly those that would come back from an intelligent Chinese speaker. That is, Tex plus the room passes what is known as the **Turing test**: to pass the Turing test is to respond to questions with all the signs of intelligence and thought distinctive of thinkers like us.

But at this point many people's intuition that we are dealing with something that does not understand Chinese starts to fade. It is still, of course, true that Tex does not understand Chinese, but an awful lot of processing is being done by Tex *plus* the book, and many argue that it is enough for the system to count as understanding Chinese.

For our part, though, we think that even after this embellishment, the system does not understand Chinese. In order to count as understanding

Chinese, we need, amongst other things, the kinds of abilities distinctive of understanding what Chinese sentences and their parts *stand for*. The system does not even understand what the word 'book' stands for, because it cannot respond in the appropriate way to a book. It only ever responds to *sentences*, not to what they or their parts stand for.

We can put this point in terms of the distinction between semantics and syntax. Syntax has to do with questions of grammatical propriety, sentential structure, whether a word is a verb or a noun, and so on. Semantics has to do with the interpretations that attach to words and sentences; it relates to what words and sentences mean. It is in virtue of having a semantics that words and sentences in a language can serve to make claims about how things are, ask questions, issue commands and so on. Understanding a language is a matter of mastering its semantics. Tex has mastered the semantics of English and in fact uses this mastery to follow the book's instructions. But as far as Chinese is concerned, all Tex has are certain syntactic abilities. He can match up Chinese characters on a screen with those in the book and type them into a keyboard, but he has no idea what they stand for or mean. In this terminology, the question we face at this stage is whether, in the embellished example, the system of Tex plus the book has a grasp of the semantics of Chinese. And our claim is that the system does not, for it does not understand what Chinese words stand for. This is manifested in the fact that it cannot respond appropriately to what Chinese words stand for. The system can only respond to words as they appear on a computer screen, not to what they stand for.

A further embellishment

We could, of course, further embellish the example. Imagine a robot which sends information about the inputs through its eyes and ears and surfaces by radio in digital form to the Chinese room. Tex is still inside, and responds to these by writing them down, sifting through the book, and following its directions as before. After much (*even* quicker than before!) calculation, he goes to the console and types a response which is relayed to the robot. This makes the robot move through and respond to its environment much as we do. We can suppose that the book is detailed enough to constitute a program for the mind of a Cantonese-speaking adult woman. Thus, the robot will interact with its environment, including answering questions, in the very same way she would.

A system that does understand Chinese

But now the intuition that we are dealing with something that does not understand Chinese has faded completely, or so it seems to us. It is still the case that Tex does not understand Chinese, but there is an entity – call her Lin – that does understand Chinese. Lin is composed of the book that contains the all-important program, parts of Tex's brain, the robot and the radio. Lin, whose robot body is entering Kowloon, might believe she was entering Kowloon. Tex, in a laboratory in Dallas, might not even know that Kowloon exists, let alone believe he or anyone else is entering it.

We should say something quickly about a response a Searle-like figure might make to replies like this. He might think that it depends crucially on there being a system bigger than Tex that does much of the work. This, he thinks, is what makes it seem plausible that the system is distinct from Tex, and can thus understand something that Tex doesn't. He might ask us instead to imagine that Tex memorizes the book, and stores all the changing data in his head. We are supposed to think that in this case there is no plausible distinctness between Tex and Lin, so if one fails to understand Chinese, so does the other. Tex doesn't understand Chinese, so nor does Lin.

We do not think that this variation makes an important difference. What we would have here is two entities who share a brain. The idea that distinct individuals might share a brain is not enormously different from that which we have grown used to in discussions of multiple personality disorder. There is some difference here; Lin relies on Tex to do the calculations that constitute her mental states (so if he gets bored or ill, it's very bad news for her indeed).

A computer analogy

It may seem puzzling that Tex does all these calculations without knowing what it is he is doing. But in fact something like this is commonplace in computer science. When one computer does calculations that emulate the behaviour of a different machine, the emulated computer is said to be a **virtual machine.** You may have seen an Apple Macintosh computer which has a window which has the look and feel of a machine running Microsoft Windows. In fact, this is done by the Macintosh operating system directing that calculations be done at the binary level that emulate the behaviour of the Intel chip on which Windows runs. If you ask the Macintosh operating system what menus appear in its windows, it will be able to tell you. But if you ask it what menus appear in the Microsoft Windows lookalike window that it is supporting by emulating the Intel chip, it won't be able to tell you. It doesn't have information about that process at that level of abstraction. If you interrogate Windows, however, it can tell you about its windows. This is roughly analogous to asking Tex about Kowloon directly, and drawing a blank, but getting an answer when you interrogate Lin.

In sum, the Chinese room example starts out as one where both intuition and any plausible functionalism agree that there is no understanding of Chinese. We can add to the example to get one where plausible versions of functionalism will have to say that there is understanding of Chinese (by the relevant entity, not Tex), but in doing so we turn the example into one where intuition also says that there is understanding of Chinese.

Blockhead

Input–output functionalism

We noted in chapter 5 that the most popular version of empirical functionalism is exposed to the charge of chauvinism. It insists on an excessive degree of internal similarity to us before something counts as having a mind; beings might fail to have minds by virtue of having internal processors which are better than ours! Would it be right to take the extra step of holding that *all* that matters for having a mind is being such as to ensure the right connexion between external inputs and outputs? Something is an amplifier if it is such as to secure the right relationship between inputs and correspondingly bigger outputs, no matter how the job is done internally. What is done, not how it is done, is what counts. Should we say the same about the mind? Such a position can insist on specifying the inputs and outputs in arm's length terms, as is done in common-sense functionalism, and that what goes on inside matters to the extent that the job of appropriately mediating between the environmental inputs and behavioural outputs must be done by what is inside. But that would be the extent of its constraints. Such a view might be called **input–output** or **stimulus–response functionalism.** It takes on board what is right about behaviourism – that behaviour in situations is crucial – but remedies at least part of what is wrong with it. Mental states are internal, causally efficacious states, *pace* behaviourism, but internal states that can be characterized fully as far as their psychological nature is concerned in terms of the behaviour that they do and would typically produce, or do or would produce if linked up in some natural way to the body. Input–output functionalism can be distinguished from **supervenient behaviourism** by the fact that the input–output functionalist insists that (most) of the states causally responsible for the behavioural profile must be internal. Suppose that Jane's normal-seeming behavioural profile is caused by puppeteers acting at a distance. The supervenient behaviourist might think she had mental states like ours; the input–output functionalist would not.

Input–output functionalism is false. A now famous example due to Ned Block shows that the way the job is done does matter. There are substantial internal constraints on being a thinker. The remainder of this chapter is concerned with describing his example – the Blockhead example – and what is to be learnt from it. We approach the example, following his lead, via some remarks about chess.

Good chess versus being good at chess

Copycat chess

It is possible to play good chess without being any good at chess. Do what expert chess players tell you to do! Your good chess will then be a sign of the ability to follow instructions and perhaps the ability to identify chess experts, but not of ability at chess. Alternatively, we can imagine that instead of turning to the experts for advice you turn to a chart or ***look-up tree*** prepared by experts. Here is how the table could be constructed for when you are playing Black. At the beginning of a game there are only finitely many moves allowed by the rules. The chess experts nominate the best response to each possible move. We obtain a little table like figure 7.1: the boxes represent the possible opening moves by White, and the circles the responses to each nominated by the experts.

Now for each response by you as Black to White's opening move, there will be only finitely many legal responses by White. So we can extend the table by asking the experts to nominate their best response to each possible response by White (figure 7.2). Again, the boxes represent White's possible moves and the circles the responses to each as nominated by the experts.

It is easy to see that this sort of diagram can ***in principle*** be extended as long as you like and that anyone who owned such a diagram – perhaps prepared by Grand Masters over thousands of years – would be able to play very good chess without necessarily being any good at chess.

Anyone playing chess with a diagram like this is very vulnerable to changes in the rules: change one rule, and the diagram can become completely useless. But we can imagine tree diagrams with various possible rule changes allowed for at any stage. Each row of boxes would be supplemented with additional boxes representing the possible moves under various rule

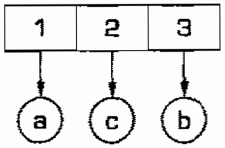

Figure 7.1 Look-up tree for the start of a chess game. The boxes represent the possible opening moves by White; the circles the responses to each nominated by experts.

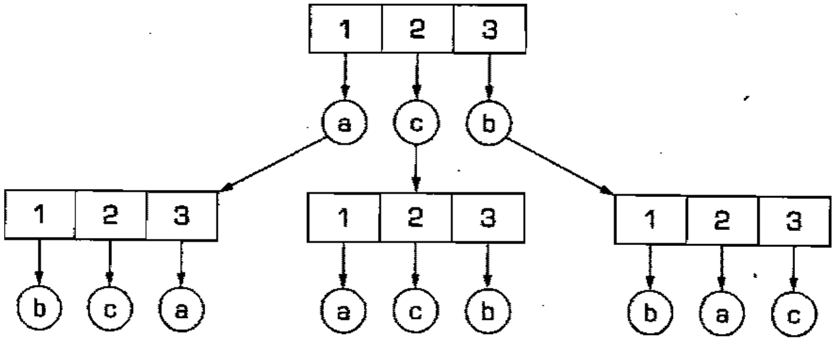

Figure 7.2 A more extended look-up tree. Again, the boxes represent moves by White; the circles represent possible responses by Black as nominated by experts.

changes, and below each box would be the circle representing the best response for that move, given that rule change, according to the experts. This would make an already *huge* tree even bigger but does not introduce any new point of principle. In practice, of course, there is an insuperable problem with this plan for playing good chess. At each stage of a game of chess there are a large number of legal moves, and for each of these legal moves there are many legal responses. Writing out the look-up tree would in consequence involve what is known as a **combinatorial explosion**. Giving more than a line or two of the tree would require more distinct states than there are particles in the universe.

The game of life

Look-up trees for life

We are now ready to describe what has come to be known as the **Blockhead** example. At any point in a game of chess, there are only finitely many legal moves and countermoves. It is this fact that makes the chess look-up tree just described possible in principle. Likewise, at any point in a creature's life there are only finitely many discriminably distinct possible inputs and outputs at its periphery. Indeed, given quantum theory, there are probably only finitely many nomologically possible inputs and outputs; but in any case we know that there is a limit to how finely we distinguish different impacts on our surfaces, and to how many different movements and responses our bodies can make. This means that there could be in principle a 'game of life' look-up tree written for any one of us – for Jones, say.

We list all the different possible combinations of pressure, light, gravity and so on impacting at the surface of Jones's body at the first moment of his life. This would be the first row of the game of life look-up tree modelled on Jones. It would correspond to the row of boxes in the chess tree above. The second row would give the behavioural response that Jones would make (in a wide sense that includes any relevant peripheral change) to each possible input. It corresponds to the first row of circles in the chess tree. The third row would give all the possible peripheral inputs for each of the various behavioural responses, and the fourth row would give Jones's behavioural responses to each of these. And so on and so forth for the whole of Jones's life. Of course how long Jones lives depends in large part on which of the various possible inputs actually come about, but we may suppose that a fail-safe strategy is employed – the look-up tree covers, say, 150 years' worth of possible inputs.

Blockhead

Jones's Blockhead twin is then defined as a creature that is superficially like Jones but has inside it a chip on which Jones's game of life look-up tree is inscribed, and this chip controls Jones's Blockhead twin's every response to the environment. In the same way we can define a Blockhead twin for each and every one of us. The objection to input–output functionalism can now be stated very simply. It is that (a) Jones and Jones's Blockhead twin behave exactly alike, not only in how they respond to each and every situation, but in how they would respond to each and every possible situation; (b) Jones's Blockhead twin is not a puppet – the connexion between its inputs and outputs is largely a consequence of how it is, not of how some puppeteer is; and yet (c) though Jones is, we may suppose, intelligent and has a normal psychology, his Blockhead twin is no more intelligent than a toaster (as Block puts it) and has no mental life at all. It really is the kind of automaton that dualists (wrongly) hold that physicalism would reduce us to.

Blockhead's challenge to us all

What is particularly interesting about Blockhead is that it tackles input–output functionalism on its favoured ground. Many find input–output functionalism implausible when applied to mental states in general, but it is at its most appealing applied to intelligence. There is no special 'feel' associated with intelligence, and it is intelligence that intuitively connects most closely with behavioural performance: the be-all and end-all of intelligence does seem to be certain capacities to deliver answers to problems set by questioners or by the environment. Yet the clearest intuition about Blockhead is precisely that it completely lacks intelligence and understanding. Indeed, given the intuitive appeal of an essentially behavioural approach to intelligence – while insisting, of course, that traditional behaviourism was wrong to refuse to see explanations in terms of intelligence as genuinely causal ones proceeding by appeal to internal nature – we should all try and say something sensible about why Blockhead is not intelligent. It is not enough to say 'It is fortunate I am not an input–output functionalist', or to say 'I hereby renounce input–output functionalism'. We all need to say *why* Blockhead is not intelligent.

Before we give our answer to this question, we note what seem to us to be some wrong turns. You might say that the reason Blockhead is not intelligent is that everything it does or would do is determined in advance. It thus lacks the flexibility that is part of being intelligent and being rational. But of course if determinism is true, everything anyone ever does is determined from the very beginning of time. Some have inferred from this that determinism is incompatible with intelligence and especially rational decision making, and have accordingly taken comfort in the fact that modern quantum theory is indeterministic. But this seems to us a pretty desperate position. It is hard to see how throwing in some random fluctuations makes what would otherwise be irrational, rational.

Some wrong turns

You might object that the look-up tree could never be written down because it requires knowing all the possibilities for inputs at any given time and all the outputs that someone – Jones, as we imagined – would make to each and every input, and such knowledge is impossible. But what matters for the argument is that the story that the look-up tree tells exists, not whether we could know it. For each and everyone of us there is a huge story about what we would do in response to each and every possible input and sequence of inputs, and so we can make sense of the idea that the story is written on a chip inside a Blockhead.

You might object that we *cannot* make sense of the idea that the story in the form of a huge look-up tree is contained inside Blockhead. The look-up tree could not exist because it would involve a combinatorial explosion. As we noted above, the look-up tree for a short game of chess, let alone the game of life, would take more particles than there are in the whole universe. This reply seems to us to misunderstand the role of thought-experiments.

The fact that Blockhead is practically, and perhaps even nomologically, impossible seems to us no more to the point than the fact that Twin Earth is practically, and perhaps also nomologically, impossible. The point of a thought experiment is to test a conceptual claim, typically a claim about the relation between two concepts. In the case of Twin Earth, we test the hypothesis that being watery and being water necessarily go together. We come up with the answer that they do not necessarily go together by making clear sense of the possibility – Twin Earth – where water is not watery, and what is watery is not water. In the case of Blockhead we test the hypothesis that being behaviourally exactly alike someone intelligent is sufficient for being intelligent, and come up with the answer that it is not, by describing a possibility we understand and comprehend (while realizing that it is in practice quite impossible) – a Blockhead twin of an intelligent Jones – where what is behaviourally exactly alike someone intelligent has no intelligence (and indeed no thoughts) at all.

Finally, you might object that though it is missing the point to complain that the Blockhead example is impossible either in practice or perhaps even nomologically, it is right to be suspicious of intuitions about cases *that* far removed from what is possible in any but the most abstract sense. Perhaps, in particular, we should resist the intuition that Jones's Blockhead twin lacks intelligence. The trouble with this objection is that Blockhead is *so* like all the cases where we feel that someone lacks understanding: someone who cannot play chess except by asking an expert what to do at every stage is someone who does not understand the game, and someone who cannot give you the square root of a number other than by looking up a table of square roots is someone who does not fully understand what a square root is. The intuition that Blockhead lacks intelligence is simply a natural extension of what we learn from these simple and familiar cases. Moreover we can give a reason why Blockhead lacks understanding and intelligence – a reason that, we will argue, makes sense of our strong intuition that Blockhead is deficient, and so explains and justifies the intuition.

Why Blockhead is not a thinker

A message of much recent philosophy has been the importance of *causal connections* of the right kind. You do not count as seeing something unless your perceptual state is caused by that thing. Part of what justifies thinking of an object – the chair in front of me or the White House – as a persisting material thing is the way early states of the object are causally responsible for later states of it. The identity through time of a person is in part a matter of their psychology being causally connected over time. Likewise, causal connections of the right kind are central to being intelligent, to rationality and to belief.

Rationality and causal history

It is not irrational *per se* to believe that the Earth is flat. It is irrational *given* what else you believe and given your history. Rationality is in part a matter of your beliefs evolving in the right way from your earlier beliefs and sensory data. We all ought to believe that the Earth is round (or oblate, to be more precise) because that is the right belief to have caused in us by our pasts. Likewise, being intelligent centrally involves having trains of thought that evolve in the right way. Later thoughts have to be caused in the right way by earlier ones. If a brain scientist inserts a probe into your brain that causes the crucial thought that enables you to announce the proof of Goldbach's Conjecture, this is not a sign of your intelligence or rationality. It is either a fluke or a sign of the intelligence of the brain scientist, depending on the causal origins of her action in inserting the probe. Moreover, it is part of being a belief of a certain kind that it tends to have certain results. Part of what makes something the belief that if *P* then *Q*, is that combined with the belief that *P*, it tends to cause the belief that *Q*. (We enlarge on the importance of tending to evolve rationally to being a belief when we discuss the intentional stance in chapter 9.)

Simple input–output devices exhibit massive causal dependencies between early and late stages. The state of a sundial or an amplifier or a carburettor that is responsible for its capacity to generate the appropriate outputs on Monday is typically a major causal factor in its capacity to do the job on Tuesday. The situation with much more complex structures like human beings is correspondingly more complex. How we respond to stimuli on Tuesday depends on all sorts of factors in addition to how we are on Monday, including what has impacted on us between the two days and what we have thought about in the interim. This is part of what confers on us the flexibility of response that makes us intelligent. Nevertheless, causal dependencies between earlier and later thoughts are crucial. It is just that how we respond in the future depends on a much more diverse range of factors than simply how we are in the past – what we have thought about and what has happened to us in the interim also enter the equation.

The trouble with devices that work by look-up tree is that they lack the appropriate causal dependencies. The state that governs the responses to inputs early on plays the wrong kind of role in causing the state that governs the responses later on. This is because, for the most part, the Blockhead is static. It is mostly written down in advance, and the only thing that varies is which node is active.

Blockhead's causal peculiarity

We will now explain the idea of an active node. We will call the various sets of pre-recorded possible inputs together with appropriate outputs *nodes*. At any given time, a Blockhead can be said to have a certain node that is *active*. The active node is the one that will be searched until the input that has been given to the Blockhead is found, and the pre-recorded output produced. Let's suppose that the Blockhead partly represented in figure 7.3 has node γ currently active. So if we give it input 1 we will get output *b*, and if we give it input 3 we will get output *c* and so on.

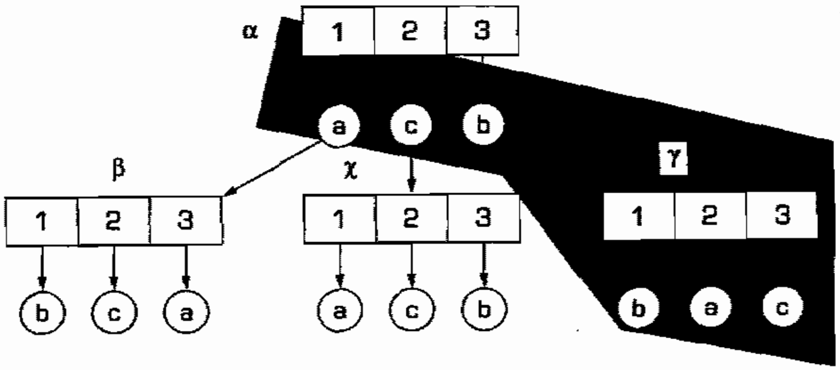

Figure 7.3 The Blockhead example – a look-up tree that simulates living an intelligent life. What actually happens lies within the shaded area.

Now *the fact that it is node γ that is currently active* does indeed depend on the past nature of the Blockhead. Indeed, it depends on the nature of node α together with the fact that node α received input 3. But that is the only thing about node γ that depends on the nature of node α, or indeed anything else about what has happened to it since its initial creation. Nothing about what possible inputs are encoded in node γ or what outputs are recorded against them depends on the history of the Blockhead.

So at the time when α is active, the later state γ that will govern the later responses already exists. The input–output profile of the look-up tree at any given time – any complete row of squares followed by circles – does not generate the profile of any particular node at any later time. If on Monday you make a sundial or Frankenstein makes a person, you do not need to do anything extra to handle how what you have made will respond to various inputs on, say, Tuesday. What you made on Monday plus what happens in the meantime does that for you, and this is crucial to the way the Tuesday responses depend on the Monday ones.

So the only causal dependency manifested in the look-up tree is in which node of the tree's input–output profile at some time the active input and output are at that time. That does depend, not on the nature of the nodes at earlier times, but on which nodes of those earlier rows were then active, and what input was received. It is like a recipe for roast duck that tells you at the end to go to another recipe for the sauce. Although what is in the first recipe plays a role in what you do subsequently, the content of the second recipe is, we may suppose, quite independent of the content of the first. But thinking is not like that; the content of what we think at a time typically depends in part on the content of what we thought at various earlier times in rich and complex ways, and that is crucial for it to count as thought and as rational thought.

In sum, Blockhead's input–output profile at any given time does not depend in the right way on its input–output profiles at earlier times for Blockhead to count as a thinker, or even as something displaying rationality and intelligence. The input–output nature of the node that controls Blockhead's behavioural response at time *t* is not caused by the input–output nature of what controls Blockhead's behavioural response at any earlier time *t* – *n*. The overall input–output nature at time *t* depends on the past states only insofar as it is determined by which pre-existing node is active. Figure 7.3 helps make the point. In the diagram, suppose that the shaded area represents what actually happens. The point is that as you progress through the shaded region you are ***not*** progressing through nodes whose nature depends on the nature of earlier states in the shaded region, except in the minimal sense in which being active counts as part of their nature.

Common-sense functionalism and Blockhead

The argument to the conclusion that Blockhead is not a thinker rests on a constraint on belief and intelligence that is supposed to be part of folklore. The way belief evolves over time, the importance for rationality of belief evolving causally in the right way, the fact that what is believed depends on what was believed and what happens to a subject are plausibly common knowledge – implicit or explicit – and part of our ordinary conception of belief. This means that Blockhead is not an objection to common-sense functionalism. Blockhead shows that input–output functionalism is false. How things are inside matters for our mental nature over and above how our insides manifest themselves in determining our environmental input–output connexions. But Blockhead does not show that common-sense functionalism is false. Indeed, we could have seen this straight off. It is ***intuition*** that delivers the answer that Blockhead has not a thought in its head. We did not describe an experiment that shows that Blockhead is unintelligent. We followed Block in supposing that once the case was described, the answer was intuitively clear – and common-sense functionalism is the version of functionalism most concerned to honour clear intuitions about the mind.

The Zombie Objection

Zombies invade the physicalist paradise

A final objection to functionalism – and indeed to any kind of physicalism – is the so-called zombie objection. Versions of this objection have been around for at least forty years, but it has become especially prominent in recent years and calls for separate treatment. The rough thought is this: we can conceive of a creature physically just like us, but which lacks those mental experiences that have distinctive 'feels' like pain and hunger, or **qualia** in the philosophers' jargon ('qualia' is a technical term for the phenomenal qualities of conscious experience, the qualities that make it the case that there is something it is like to have the experience; see the next chapter for a more detailed explanation). Let's call them **zombies**, for they walk and talk like real people, but lack qualia or conscious experience. From this we conclude that it is possible that there could be such a creature. But then there must be something about the actual world other than its physical make-up which gives us qualia: for this possible world we have imagined is one which has exactly our physical make-up, but lacks qualia. The thing that it lacks can't be physical, since it is exactly like the actual world in physical respects. So the thing in the actual world that makes it true that we have qualia must be non-physical – thus physicalism is false.

Let's be a bit more precise, and lay the argument out:

Premise (1). We can conceive of a world physically exactly like ours, but which lacks any other features ours may or may not have (i.e. it is a minimal physical duplicate of ours), in which the people in it lack qualia and consciousness – they are zombies.

Premise (2). Conceivability is a good guide to possibility.

Premise (3). We are not zombies.

Intermediate conclusion (4): Zombies are possible (from (1) and (2)).

Intermediate conclusion: So there is a minimal physical duplicate of the actual world that is mentally different from the actual world, since it contains only zombies whereas the actual world does not (from (3) and (4)).

Conclusion: Physicalism is false, from the definition of physicalism in chapter 1.

The argument is logically valid. So if all its premises are true, then so is its conclusion. Any physicalist must, therefore, deny one or more of its premises. Exactly which premise they deny, however, varies according to the kind of physicalism involved.

Analytic functionalism needs to deny the first premise. For analytic functionalism says that it is a matter of the meaning of mental state terms that you have the relevant mental states whenever you have the right functional roles being played. And if the relevant functional roles are played actually by physical stuff, then those roles will be played in any minimal physical duplicate of the actual world. Suppose that an analytic functionalist has worked through her theory of mind, and knows what roles have to be played for qualia to exist. She ought not be able to conceive of zombies. It is ***a priori*** that zombies are impossible. For knowledge of the roles, together with knowledge that a physical set up which plays them exists, logically entails that qualia exist. To conceive of zombies is to conceive of things that have what is sufficient for qualia (having the right roles played), and yet lack qualia. And this is to conceive of a straightforward contradiction. In some good sense of conceive, one cannot conceive of the ***a priori*** impossible.

ıalytic functionalism
d ideal
nceivability

This is a real problem for analytic functionalism, for it seems that we really can conceive of zombies – and yet the analytic functionalist says they are ruled out by our grasp of the meaning of mental state terms. It is intuitively fine to think it might be true that zombies are impossible, but perhaps only if this is a substantial fact that does not simply fall out of the meaning of mental state terms. Many think that if zombies are impossible it does not seem to be a merely semantic fact, but rather a metaphysical one. There are, however, things that analytic functionalism might do to sweeten the pill. Sometimes we think we can conceive of something that is impossible. Perhaps you were asked once, in maths class, to find out at what point a parabola crossed the *y* axis. You took very seriously that it was at $y = 1$ and $y = 2$. You not only conceived of that possibility, but you thought it actually true. But after some calculation you found, no, it was at $y = 2$ and $y = 3$. But of course, once we define a parabola by a quadratic equation, it is a logical necessity that it intersects the *y* axis (or not) where it does. Your original conception was incoherent: it was a logical impossibility. Yet you had it none the less. Exactly what to say about this case is controversial. Everyone agrees, however, that there is some important distinction between what you can ***ideally*** conceive – that is, when all the logical and semantic truths are before your mind and you are rational – and whatever is happening when you 'conceive' of the roots of a quadratic equation being different from what they, of necessity, are. The disagreements are about what is going on in the unideal case, and we can set that aside. For perhaps the analytic functionalist should say that they doubt that we can ***ideally*** conceive of zombies. If all the facts about the functional roles were before your mind, and you could see how the physical states must play those roles, you could not conceive of zombies. Our apparent ability to conceive of zombies is on a par with imagining that mathematical truths are false.

This is a powerful reply, and one of the authors is very attracted by it. However, it remains a little mysterious how all the extra clarity and ideal rationality are supposed to do their work. Certainly *if* it follows from the meaning of mental state terms that zombies are impossible, then we ideally can't conceive of them. But one might take a strong intuition about their conceivability to be evidence that we have got the theory of the meaning of mental state terms wrong. So for the reply to work, the extra clarity that we have ideally will need to make it clearer that the analytic functionalist theory of mental state terms is right – which is a punt the analytic functionalist must take. In addition it is hard to imagine exactly what form this extra clarity would take. We sort of know what the extra clarity would come from in the mathematical case, but what the analogue is in the qualia case is harder to see.

Empirical functionalism and zombies

At first glance, empirical functionalism appears to be in a better position to address the zombie challenge. For here the discovery that qualia are physical is in some sense ***a posteriori***. So, the thought runs, we might be able to conceive of zombies, for it is not ***a priori*** that they are impossible. They are impossible none the less, but this is an ***a posteriori*** impossibility of the kind we discuss in chapter 4. Thus the empirical functionalist might try to deny the second premise of the zombie argument. For they think we can conceive of the impossible – even ideally conceive of the impossible. Thus conceivability is no guide to impossibility, and the **zombie argument** fails.

We think, however, that this tempting reply fails. It fails because it ignores some subtle distinctions within empirical functionalism. On one kind of empirical functionalism the view is coherent but the reply does not work. On another kind, the reply seems to work – but the coherence of the empirical functionalism itself is problematic.

Reference-fixing

Many of the versions of functionalism that we identify in the table at the end of chapter 5 use something – perhaps the folk roles – to pick out some mental natures, and then rigidify on the internal features of those entities. Now it is impossible that something be a physical duplicate of that internal nature without possessing that internal nature, on the assumption that the internal nature we discover is physical. So suppose that whatever it is that we use to reference-fix, reference fixes on some human brains. When we do some empirical work on brains we discover Neural Feature X playing the qualia role, and we conclude that *a posteriori* we have discovered that qualia are instances of Neural Feature X. If we rigidify (see chapter 4 if this is opaque to you) we will in addition conclude that qualia are *necessarily* neural feature X. Thus it is impossible to possess neural feature X without possessing qualia. Thus zombies are impossible – for by definition they lack qualia, but in virtue of being physical duplicates of us they possess neural feature X, which means they *do* possess qualia. So on these versions of functionalism zombies are impossible, but it would appear to be *a posteriori* that they are impossible. It appears not to be a conceptual truth, so the mere conceivability of zombies poses no threat. After all it was conceivable that many *a posteriori* necessities (such as water's being H_2O) were false.

But the empirical functionalist cannot rest too easily. Remember that on many of these views the folk roles were used to reference-fix on the samples whose internal nature was discovered. Grant to the empirical functionalist that 'qualia' is a term that refers to the thing that actually plays those roles – neural feature X. We still need a term for the thing we knew about before we knew any neuroscience, and that enabled us to reference-fix. Let us use 'qualic' to mean having the folk roles played. Before we knew neuroscience, we did not know that neural feature X was in us, but we knew that we were qualic, and (let us further suppose) we knew that qualia were the actual qualic things.

Now it is surely conceivable that there are possible physical duplicates that fail to be qualic (incidentally, if you think that being qualic just is having qualia, then you agree with us that empirical functionalism is a bad idea). Call them R-zombies. But if being qualic is just a matter of playing the folk roles, then it follows as a matter of the meaning of 'qualic' that nothing could be physically just like us and fail to be qualic. But this clashes with the intuition that there can be R-zombies – beings that are physically just like us but lack the feature that we used to identify neural state X. In a nutshell, while the empirical functionalist can accept a zombie intuition with respect to the features they call 'qualia', the problem arises again at the level of the reference-fixing descriptions. For them, R-zombies must be inconceivable, even though they seem perfectly conceivable.

You may recall that some versions of empirical functionalism about qualia deny that the folk roles have any job to do. They take it that we brutally discover the empirical nature of qualia. These versions seem doomed to run into one of two major problems. If they say what it is that we use

to decide which things have a qualitative nature, then a zombie objection of the kind we mention for empirical functionalism with the folk roles seems near at hand. If, on the other hand, they do not, it is completely mysterious why they think we should be examining heads rather than rocks in our search for the empirical nature of qualia.

A modification of analytic functionalism

*Perhaps the best reply proceeds something like this. The analytic functionalist should never have been entirely sure that analytic functionalism is true. It is, after all, a controversial doctrine. In any case dualism might be right – no matter how unlikely. So the analytic functionalist should think at most that if dualism is false, then analytic functionalism is true. But then the conceivability of zombies is in part the conceivability of dualism's being true. Now if the analytic functionalist holds that the truth of analytic functionalism is an *a priori* truth there is still a bit of a problem, for zombies should still be *ideally* inconceivable, since ideally we will grasp the truth of analytic functionalism.

But there is a weakening of analytic functionalism that is available. Perhaps all that is *a priori*, and follows from the meaning of our mental state terms, is that *if dualism is false* then analytic functionalism is true. Perhaps the analytic truth – the thing that we grasp by grasping the meaning of our mental state terms but not knowing empirically how the world is – could be something like:

If there are dualistic states *then*

- In the actual world the qualia are the dualistic states; and all and only the qualia in counterfactual worlds are the dualistic states.

Else

- In the actual world the qualia are the states that play the functional roles, in other worlds qualia (if any) are the states that play these roles in that world.

If this is the right analysis, and if dualism is false, then it will certainly be right to say 'zombies are impossible'. But we wouldn't know that we were entitled to say that *a priori*. For we know that we are only entitled to say that zombies are impossible if dualism is false, and even physicalists should give some credence to dualism in fact being true, however small. The idea is that the zombie intuition comes from confusing two things. The first of these is the thought that there is some chance that dualism is true – and on that supposition we would be right to claim that zombies were possible. The second of these is the straightforward possibility of zombies.

Annotated Reading

The Chinese nation example is presented in Ned Block, 'Troubles with Functionalism'. John Searle's Chinese room case has been very widely discussed (at times, with some heat). Perhaps the best place to start is John Searle, 'Minds, Brains, and Programs'. A more informal presentation, combined with replies to the many objections that have been raised, is his 'Is the Brain's Mind a Computer Program?' Among the many replies he considers are those he christens the systems reply and the robot reply. The systems reply is the first one we expounded. The reply we eventually settled on is a combination of the systems and robot replies. A good recent discussion is in chapter 6 of Jack Copeland, ***Artificial Intelligence***. The classic source for Blockhead is Ned Block, 'Psychologism and Behaviourism'. Keith Campbell's ***Body and Mind*** provides a straightforward description of the zombie argument (he uses imitation men instead of zombies). The term 'zombie' in this context may have come via Robert Kirk, 'Zombies versus Materialists'. Recent interest in the zombie objection has been stimulated by David Chalmers, ***The Conscious Mind***. A fuller version of the reply we give in the final section can be found in David Braddon-Mitchell, 'Qualia and Analytic Conditionals', and for a similar approach see John Hawthorne, 'Advice for Physicalists'.

8

PHENOMENAL QUALITIES AND CONSCIOUSNESS

Mental states and 'feels'

An itch feels different from an ache. A stabbing pain feels different from a burning one. But the belief that two is the smallest prime does not feel different from the belief that the Earth is oblate. Beliefs don't have 'feels'. Again, seeing something that looks red is a different experience from seeing something that looks green, but hoping that the drought will break is not a different experience from hoping that the cheque will not bounce. Hoping is not an experience, though it is sometimes associated with various experiences – of relief when the cheque does not bounce, of joy when the rain arrives. In what has become a common way of putting the distinction, we distinguish those psychological states for which there is ***something it is like to be in them*** from those for which the notion seems to make no sense.

Bodily sensations and perceptual experiences are prime examples of states for which there is something it is like to be in them. They have a **phenomenal feel**, a phenomenology, or, in a term sometimes used in psychology, are raw feels. Cognitive states are prime examples of states for which there is ***not*** something it is like to be in them, of states that lack a phenomenology. These terms – 'phenomenal feel', 'having a phenomenology', 'there being something it is like to be in them', 'raw feels' – are not exactly transparent. They are ways of getting you to identify the distinction we have in mind on the presumption that you are already familiar with it. If you are not already familiar from your own mental life with the distinction between mental states that have a distinctive feel and those that do not, no words of ours will help you grasp it. Our words are not intended to inform you of a distinction you were previously ignorant of, but to identify for you the distinction this chapter is concerned with.

There is debate about which states fall into which category. What about desires and emotions? Desires, particularly desires for food and sex,