6

Thought

In hitting upon the formulation "I think, therefore I am," Descartes took himself to have established not only his existence, but his nature: he is essentially a thing that thinks. Thought, that is to say, is the essence of mind. There are two aspects of thought that are of particular philosophical interest: its representation of things beyond itself, that is, its intentionality; and its movement from one representation to another in accordance with the laws of logic, that is, its rationality. But, as indicated in the previous chapter, contemporary philosophers of mind typically take the problems of qualia and consciousness to pose the most serious challenge to a materialist concept of the mind, with intentionality and rationality being more readily explicable in naturalistic terms. There is a certain irony in this view; in so far as it effectively takes sensation and feeling - capacities we seem to share with other (obviously material) animals- to be more mysterious than thought, which we (arguably) do not share with them. One would have thought it more natural to see things the other way around; indeed, most philosophers of the past have seen things the other way around. The suggestion that what we share with the beasts is scientifically puzzling, while what appears to be unique to us is merely one, relatively unproblematic material capacity among others, would have struck Plato and Aristotle, Augustine and Aquinas, Descartes, Leibniz, and Kant as odd, even perverse.

We also saw, in the previous chapter, that there is a strain in contemporary thinking that holds qualia and consciousness ultimately to be

explicable in terms of intentionality, and it was suggested that a strong case could be made for this view. But, in so far as the same strain typically takes the task of explaining intentionality itself in materialistic terms to be little more than a comparatively trivial mop-up operation, it is, arguably, misguided. As we shall see, a number of contemporary philosophers hold that the older philosophical tradition was correct, and that there are considerable difficulties involved in carrying out a naturalistic explanation of thought. In this chapter and the next we will examine recent attempts at such an explanation. This chapter will focus on attempts to account for rationality in particular; and we will see that, as with our investigation of qualia and consciousness, the investigation of rationality leads us inexorably to intentionality. Chapter 7 will then deal, at last, with that most ubiquitous of mental phenomena.

Reasons and causes

Suppose you witness Ethel crying out in pain after stubbing her toe, and then watch as she removes her shoe and examines her foot. If asked to explain the first event, you would probably say something to the effect that the damage to her body resulted in her crying out; if asked to explain the second, you would say that she wanted to determine the extent of the damage and thought that removing her shoe would be the best way to do so. In the first case, you would be pin-pointing the causes of her behavior; in the second you would be giving the reasons for it. In both cases you are giving an explanation of human behavior, but the sort of explanation is very different in each. In the first you are appealing to brute physical forces - an impact on skin and muscle tissues, together with the stimulation of nerve endings - while in the second you are appealing to what a person takes to be a rational course of action given her beliefs and desires.

This distinction between reasons for and causes of behavior is a crucial one, and raises in a vivid way the question of how human beings fit into the natural world. The role of causes seems unproblematic. The human body is, after all, a material system alongside other ones, and it is, as much as they are, governed by the causal regularities enshrined in the laws of physics. So it is not surprising that much of human behavior should be explicable in causal terms. But what about behavior that seems to involve more than this? What about behavior that results from choice, after reflection about which course of action would be best? To understand such behavior, it seems insufficient to

speak in terms of ordinary causal factors - the stimulation of nerve endings, the secretion of chemicals, the firing of neurons and the like. Reasons for the action taken are relevant also, and appear to be just different sorts of things from causal factors. To say that neural processes cause the muscles in my fingers to move as I type these sentences is true enough; but my desire to write these sentences, my belief that using a word processor would be the most efficient way of doing so, and my consequent decision to start typing are clearly just as important, and seem irreducible to the sorts of causal processes alluded to. For A to be the cause of B is one sort of relation; for A to be a reason for B is another. The first concerns the impersonal realm of meaningless material forces; the latter concerns the personal sphere of rational deliberation. It's a straightforward case of comparing apples and oranges.

The trouble is that giving a materialistic or naturalistic explanation of any phenomenon seems somehow to require fitting it into the causal network described by physical science. If the materialist picture of the world is correct, there can be no true explanation of human behavior that does not ultimately amount to a causal explanation. But are the reasons one has for an action really analyzable in terms of causes of that action, appearances notwithstanding? Many philosophers have thought so. They would argue that since the action of my typing these sentences was the result of the reason for action constituted by my beliefs and desires, there is a clear sense in which it was caused by that reason for action. Reasons are, on this view, just a species of causes. But other philosophers have, following Ludwig Wittgenstein (1889-195 I), argued that, in many cases, it is simply a conceptual confusion to treat reasons as causes of action. The smile with which I greet you is, in this view, not caused by the happiness I feel at your return from a long trip, even if the happiness was the reason for my smile; rather, the smile partially constitutes the happiness. The behavior and the happiness are not two neatly distinguishable elements related, like events as described in physical science, by some causal law. The tie between them is an intrinsic, conceptual one.

What we want to focus on, however, is not the question of whether this or that isolated reason for an action might plausibly be said to be a cause of the action, but instead on the larger question of whether the vast network of beliefs, desires, thoughts, and other propositional attitudes as a whole, which largely constitutes the mind, can plausibly be explained in terms of the network of causal processes that constitutes the brain. We noted in chapter 3 that the elements of the first network

are related by logical connections, whereas the elements of the latter are causally related. When one set of neural processes brings about another, this is at most an instance of a contingent causal regularity. But when the thought that all men are mortal and Socrates is a man brings about the thought that socrates is mortal, this is a case of logical inference, where the second thought follows of necessity. So how can the latter sort of phenomenon possibly be explained by reference to the former? How can the wholly contingent tendency of certain neural processes to trigger certain other ones account for our ability to think in accordance with the utterly inflexible laws of logic?

The computational/representational theory of thought

The answer, in the view of many contemporary philosophers of mind, lies in the digital computer. We saw in chapter 3 that one way of expanding on the generic functionalist idea that mental states are definable in terms of their characteristic causes and effects is to think of those causes and effects as the inputs, outputs, and transitional states of a computer program. The mind, in this view, is literally a complex piece of computer software implemented on the hardware of the brain. The modern theoryof computation owes much to the mathematician Alan Turing (1912-1954), whose concept of a Universal Turing Machine - an abstract specification of a mechanical device capable of implementing any algorithm - was the model for the modern computer. The view in question is thus sometimes called Turing machine functionalism.

The beauty of an algorithm is that it provides a way of carrying out a highly complex task - including such tasks as performing a difficult mathematical computation, or reasoning through a long chain of argument to a conclusion - in a series of simple steps. The steps can in fact be so simple that we often speak of carrying them out "mechanically." And what a computer does is essentially to mimic, in this mechanical way, what we do when we follow an algorithm. Your pocket calculator or computer perform a number of elementary operations, realized in nothing more than the sending of electrical signals, which collectively add up to something significant: the display of "4" following upon the inputs "2," "+," "2," and, "=," or the generation of text following upon the pressing of keys on a keyboard. Since the elementary operations are so extremely simple, it is possible to construct a machine which is capable of performing them with a very

high degree of reliability. And this means that it is possible to construct a purely material system whose operations parallel exactly the laws of logic. A suitably programmed computer can be depended upon always to display "4" following the inputs "2," "+," and "=," and always to generate "Socrates is mortal" following the inputs "All men are mortal" and "Socrates is a man."

If an artificial device can do this, why not a brain? Why can't we suppose that neural processes are as capable of implementing algorithms as are computers? Indeed, perhaps this is exactly what human thought, including the most abstract and rigorous mathematical and logical reasoning, really is: the implementation of a set of algorithms constituting a program. And if so, the way would be opened to fitting the sphere of reasons for action, and reasoning in general, into the sphere of physical causation. Just as the implementation of a computer program is ultimately reducible to the network of causes and effects instantiated in a piece of computer hardware, so too would the implementation of the program that is the human mind be reducible to the network of neuronal firing patterns constituting the brain. The capacity of the brain, considered as a purely material system governed bythe same laws of physics that govern everything else in the universe, to generate patterns of thought that correspond to the laws of logic would be no more mysterious in principle than the capacity of a calculator reliably to function in accordance with the laws of arithmetic.

In a computer there are identifiable symbols - numerals like "2" and "4," and the signs "+" and "=" and so forth - that correlate with the numbers and functions of a mathematical computation. Is there anything analogous in the case of the computer that is the brain? Many philosophers have argued that there is, in the form of sentences. In their view, a particular mental state, such as the belief that Socrates is a man, is to be understood as a relation between the person having the belief and a sentence that has the meaning that Socrates is a man. where is this sentence, though? surely it can't be in the brain itself - there is nothing in the brain that looks like the sentence "socrates is a man." And what language is this sentence written in? surely not English, since lots of people who do not speak English have the belief that Socrates is a man.

It is a mistake, however, to suppose that a sentence having the meaning that Socrates is a man has to look like the sentence "Socrates is a man." After all, the sentence "socrates is a man" could be handwritten instead of typed on paper, and remain the same sentence despite the difference in appearance. Moreover the sentence could be

spoken, existing only as sound-waves rather than splotches of ink on paper; if spoken into a tape recorder, it would exist as a pattern on recording tape. So why couldn't it exist as a neuronal firing pattern in the brain? Why couldn't there literally be "sentences in the head," as some theorists have put it?

If there are such sentences they would indeed not plausibly be sentences of English - or Spanish, Chinese, German, or any other natural language. But they could well be sentences of some other, universal language - a "language of thought" common to all human beings, one we all think in unconsciously, and the sentences of which get manifested in our conscious thinking, speaking, and writing as translations (as it were) into sentences of English, Spanish, Chinese, German, and all the rest. Philosophers who take the view that there is such a language of thought often refer to it as Mentalese, and since the overall theory of which the Mentalese hypothesis is a part is one that takes thought to be computation of a sort analogous to the computation performed by modern digital computers, where this computation involves transitions between states directed on to sentential representations in a language of thought, the theory is often referred to as the conputational/representational theory of thought or CRTT (in the words of Jerry Fodor, the theory's best-known advocate). Its defenders claim that, whatever else one thinks of this theory it shows that there is, in principle, no problem in explaining our capacity for rational thought in purely materialistic terms.

The argument from reason

There are, however, a number of serious objections to this proposal. Consider first the implications of taking mental states to be states of a computer program whose causal efificacy derives entirely from their implementation in electrochemical processes in the brain. When you type"2," "+," "2," and "=" on the keyboard of an electronic calculator, various electrical signals are sent through the device which ultimately cause the symbol "4" to appear on the dlsplay screen. But that that symbol signifies to us the number 4, and that the other symbols signify the number 2, the function of addition, and the relation of being equal, plays no role whatsoever in the causal process. If we decided to change the meanings of these symbols - for instance, by using the sequence "2 + 2 =" to mean "Please display the message that it is raining" and the symbol "4" to mean "it is raining" - this would have no effect on how the device operates. Nor would it have any effect if we all forgot the

meaning of the symbols, and came to regard calculators merely as toys that displayed different shapes whenever one pressed their keys. The meanings of the symbols are, in short, completely irrelevant to their causal efficacy, for they would have the same causal properties whatever meanings they had, or even if they had no meanings.

If this is true of the symbols processed by a calculator it would be true also of the symbols "processed" by the brain - it would be true, that is to say, of the contents of our thoughts as they are characterized by the CRTT. If your thought that "socrates is a man" is identical with a neural process instantiating a sentence in Mentalese which has the meaning or content that Socrates is a man, then that meaning per se plays absolutely no role in causing whatever events the neural process, and thus the thought, causes. The causal properties of the neural process/ thought would be just as they are even if it had instead the meaning that "it is raining," or even if it had no meaning at all. And that entails that the fact that your thought has the content that "socrates is a man" plays absolutely no role whatsoever in causing you, for example, to say or write the sentence "socrates is a man." You would have written or uttered the same sentence even if your thought had been about the rain or even if it had had no meaning at all. The electrochemical properties of the neural process implementing the thought are all that matter to its causal efficacy, just as the electronic properties of the symbols in a calculator are all that matter to their causal efficacy.

What this seems to mean is that distinctively mental properties turn out in the materialistic CRTT to be no less epiphenomenal than they do with property dualism. Nor is the CRTT the only materialist theory to have this consequence; indeed, any theory that takes mental states to have whatever causal efficacy they have only because of their identity with or supervenience upon physical states seems destined to have the same result: the physical properties of such states end up doing all the causal work, with the mental properties being an irrelevant, epiphenomenal extra. Epiphenomenalism would thus appear to threaten materialist theories no less than it does dualist ones - in which case the claim of materialist theories to be better able than dualist ones to account for the causal relations between mind and body seems to dissolve.

The problem, however, seems especially poignant for the CRTT, given its claim to provide a materialistic explanation of our capacity for rational thought. If the content or meaning of thoughts has, in the CRTT, no causal influence on behavior, neither does it have any causal influence on other thoughts. That your thoughts have the content that

Socrates is a man and that all men are mortal can have no influence whatever on producing the thought that Socrates is mortal, for that last thought would have been caused by the others even if those others had instead had the content that Fido is a dog and all fish have fins, or even if they had no content or meaning at all. The electrochemical properties of the neural processes with which the thoughts are associated are entirely sufficient to bring about whatever effects they do bring about. The meaning or content of the thoughts is irrelevant.

That this result is as counter-intuitive as it is is bad enough, but the problem goes deeper. It is only in virtue of the meaning or content of thoughts that they can serve as a rational justification for other thoughts: your thoughts that Socrates is a man and that all men are mortal are a rational justification for believing that Socrates is mortal only because they have the meaning they do, and they would not serve as a rational justification for the latter thought if they meant instead that Fido is a dog, etc. Yet if the meaning or content of a particular thought plays absolutely no role in bringing about any other thought, it would seem to follow that it can provide no rational justification for any other thought. You'd have exactly the same beliefs you have now whatever the content had been of the further beliefs you appeal to in justifying them. In that case, however, your beliefs would seem to have no rational justification at all. But surely this cannot be right – surely you do have a rational justification for at least many of your beliefs. Yet the CRTT, it seems, cannot account for this - ironically enough, given that its very rationale was to account for our capacity for rational thought. Even worse, advocates of the CRTT obviously think they have a rational justification for their own belief in the CRTT; but if the theory is correct, it would seem that they can't! The theory appears to undermine itself.

The CRTT defender might appeal to evolution as a guarantee of the reliability of our thought processes: wouldn't natural selection ensure that our brains are wired in such a way that the thoughts we generate are, for the most part, true? Wouldn't we have died out long ago if things were otherwise? One quick reply to this would be to suggest that it is question-begging: for it assumes that we can be rationally justified, in the CRTT, in believing the Darwinian evolutionary story (or believing anything else) in the first place, which is precisely what is at issue. Another reply would be to note that what natural selection tends to maximize is the capacity of an organism to survive and reproduce, and there is no reason to assume that having a true system of beliefs really is what is most conducive to survival: maybe our environment is

such that we have been able to survive and reproduce as well as we have only because we have developed a mostly false system of beliefs, a kind of elaborate fantasy world that shields us from certain truths, the knowledge of which would tend toward our destruction (perhaps because they would be too horrifying for us to bear). But there appears to be an even deeper problem. The general truth or falsity of a system of beliefs can only be affected by natural selection if that system of beliefs has, by virtue of its truth or falsity, some causal influence on behavior - that is, if the truth or falsity per se causes behavior which is either adaptive or maladaptive, and whith will tend therefore to get either selected for or selected out. But a belief's being either true or false is bound up with its having the particular content it has, and as has been suggested, there seems to be no way, in the CRTT (or perhaps in any materialistic account of thought), for the content or meaning of a thought to have any causal influence on behavior. The purely neuro physiological properties which, according the CRTT, instantiate the thought are the only ones that can have any causal relevance. So there is no way for the truth or falsity of a belief to have any efifect on behavior, and thus natural selection cannot affect in any way the general truth or falsity of a system of belief. But in that case, if the CRTT (or any purely materialistic account of thought) is true, evolution cannot account for the reliability of our thought processes.

The sort of argument described in this section is sometimes called the argument from reason, and versions of it have been presented by C. S. Lewis (189&-1963), Karl Popper (1902-1994), and, most recently, Alvin Plantinga and William Hasker. In so far as it depends on the claim that materialist theories cannot avoid epiphenomenalism any more than property dualism can - the claim, that is, that materialists cannot solve what philosophers of mind have come to refer to as the "problem of mental causation" - it rests on a premise that is bound to be controversial. But it shows, at the very least, that the suggestion that our capacity for rational thought is in principle easily explicable in naturalistic terms is far from having been demonstrated.

The Chinese room argument

Many think that this conclusion is bolstered by an important set of arguments associated with John Searle - perhaps the foremost critic of the notion that the human mind ought to be thought of as a kind of software and the brain as a kind of computer hardware. The first and most famous of these arguments involves a thought experiment that

has come to be known as the "Chinese room," and is directed at the claim that the implementation of the right sort of program – whether in a computer, a sophisticated robot, or a human being - is sufficient for genuine intelligence. Searle asks us to imagine a scenario in which he is locked in a room with a collection of Chinese symbols and a rulebook, written in English, which tells him which combination of symbols to put together in response to questions written in Chinese and slipped to him through a slot in the door. Searle doesn't speak a word of Chinese, and the rulebook doesn't tell him the meanings of the symbols he's combining - all it tells him, in effect, is that when he's given a set of symbols that look like this (where this refers to some specific set of shapes on the page), he should reply with a set of symbols that look like that (where that refers to some other set of shapes). It is possible that Searle could get so god at combining the shapes that a native Chinese speaker who is putting questions to him through the slot and is unaware of what is going on would assume that Searle really speaks Chinese.

Turing famously suggested that a way of determining whether a suitably programmed machine could be said truly to think would be to put it in a situation where a human being would have to carry on a conrversation with both the machine and another human being, and try to determine which participant in the conversation was the machine and which the other human being. If, after a sufficient period of time, the interlocutor couldn't determine which was which - if, that is to say, the machine's performance was indistinguishable from that of the human being - then, Turing suggested, the machine could be regarded as having exhibited real intelligence. The appropriate way to test for intelligence, on this view, is to see whether something behaves intelligently, and the machine will have passed what has come to be known as the "Turing test."

Searle, in his Chinese room, exhibiting behavior that is indistinguishable from that of a native Chinese speaker, has thereby passed the Turing test for understanding of Chinese. Moreover, he has done so by doing what a computer program does, namely, manipulating symbols in accordance with an algorithmic procedure to which only the symbols' physical properties (in this case their shape), and not their , meanings, are relevant: he is, in effect, "running the program" for competence in the Chinese language. Yet for all that, he still does not understand a word of Chinese, and has no inkling of what the answers he's giving out mean. (Perhaps he occasionally hears some yelling on the other side of the door and wonders whether he's just "said" some

thing insulting, or hears laughing and wonders whether he's told a joke or committed a faux pas!) But then it follows, Searle concludes, that running a program, of whatever level of complexity, cannot suffice for understanding or intelligence; for if it did suffice, then he would, simply by virtue of "running" the Chinese language program, have understood the language. So human intelligence just isn't what the CRTT says it is: it is not the implementation of a kind of computer software.

Searle considers the possible reply to this argument that even if he doesn't understand Chinese, it doesn't follow that no understanding of Chinese is present. After all, it isn't just a part of a computer, even the central processor, that runs a program, but the computer as a whole; and Searle is, in the thought experiment, part of a larger system that comprises also the rulebook, symbols, and door slot. It is this entire system which, strictly speaking, runs the Chinese language Program. So maybe the system taken as a whole understands Chinese, even if one part of it (Searle) does not. This "systems reply" (as it is known) may sound bizarre: how can a room, even one as eccentric as the Chinese room, be said to "understand" Chinese, or anything else for that matter? But if one is willing to take seriously the suggestion that intelligence consists of the running of a program in the first place, one is bound to have to swallow some unusual consequences, given the great variety of systems which could, in principle, implement a program. In any event, Searle argues that the room is not really essential to the thought experiment. We could instead imagine that he memorizes the symbols and rulebook, and responds to questions put to him by quickly recollecting what symbols to give out in response to whatever symbols are put to him. Perhaps he even memorizes the sound of each symbol as well as its shape, and, following the rulebook, can now respond verbally to whatever is said to him by uttering the appropriate sequence of (what to him sound like) noises. In this scenario, Searle himself just is the entire system - yet he still doesn't understand a word of Chinese.

Some have suggested that in this scenario - in which, we can suppose, Searle interacts directly with other speakers and with the external world - he inevitably would pick up on the meanings of the Chinese words he's uttering. If a certain sequence of sounds tends to be uttered only when it is raining, he's bound to be able to infer that it means "it's raining"; if another sequence tends to be uttered when cheeseburgers are in the vicinity, he might conclude that it means "cheeseburger," and so forth. Whether such causal interaction with the

world would suffice to generate a grasp of meaning is something we'll explore in the next chapter. But, as Searle notes, even if such an account is correct, the reply to his argument just sketched essentially concedes its main point, namely, that running a program is by itself insufficient for understanding.

There is a way to argue that in Searle's revised scenario, genuine understanding of Chinese would, for all Searle has shown, exist even in the absence of causal interaction with the world. Consider the fact that computers often run a number of programs simultaneously; for example, you might surf the Internet, and thus be running your web browser, while also playing a video game and typing a paper with your word processing software. Yet though the same machine is running all three programs, none of the programs necessarily has any influence on any of the others. Your word processing has no effect on your score in the game, and your score has no impact on which websites you visit. You might say that none of the programs "knows" what the others are doing. But maybe something similar is happening with Searle: his conscious understanding of English might be identical to his running a certain program (the program for English competence), while at the same time, by virtue of his following the rules in the rulebook and implementing the program for Chinese understanding, there is a second stream of consciousness that is consciously aware of speaking and understanding Chinese, even if the English-speaking program isn't. Since they are different programs, neither has any access to what is going on with the other one, any more than your word processor "knows" what your web browser is up to; but that doesn't mean that each one isn't aware of what is going on within itself. The result would be something like Multiple Personality Disorder: by virtue of his running both the English- and Chinese-speaking Programs, more than one mind has taken up residence in Searle's body, though Searle is aware only of the thoughts of the first. If this is possible, then the fact that Searle's English-speaking stream of consciousness wouldn't be aware of understanding Chinese would nevertheless be consistent with there being some stream of consciousness within him that does understand it, and if that possibility hasn't been ruled out, the computational picture of the mind hasn't been refuted.

Other defenders of the CRTT have suggested that the replies to Searle's argument just surveyed fail to get at its main problem, which is that it is really directed at a straw man, Fodor, ~~in particular~~, has argued that it is a mistake to view the computational/representational approach to the mind as a theory of understanding in the first place.

Advocates of that approach do not hold - or at least need not hold, and should not hold - that it gives an account of meaning or intentionality: it has nothing to say about how symbols, Chinese or otherwise, come , to have anycontent, or about how we come to understand that content. Rather it is merely a theory about rationality, about our ability to go from one thought to another in accordance with the laws of logic; and what it holds, as we've seen, is that we are able to do this because our thought processes are computational processes implemented in the hardware of the brain. Nothing in Searle's argument undermines this claim: he is, by virtue of "running" the Chinese language Program, genuinely engaging in rational thought, even if he is unable to understand the contents of the thoughts he's having. Of course, this doesn't show how the CRTT can get around the other objection we've looked at - the argument from reason - but it does seem to show that the Chinese room argument cannot provide compelling, further, independent grounds for rejecting the CRTT.

The mind dependence of computation

The Chinese room argument seems, at best, inconclusive. But Searle has other arrows in his quiver. The claim of computationalism is that the human mind is identical to a computer program, a piece of software implemented in the brain. The brain, that is to say, is on this view literally a kind of computer. But by virtue of what, exactly, does something count as a computer in the first place? Consider the computer sitting on your desk. You use it to surf the Internet and do word processing, and part of what this involves is the generation of text and images on the computer screen in response to inputs tyed on the keypad. As we've noted earlier; the words and images appearing on the screen are intrinsically just meaningless patterns, shapes, and colors: it is we who give them whatever meaning they have; the same images could, in principle, have come into existence accidentally, and been associated with no meaning whatever. But Searle argues that the same thing is true of the electrical impulses produced by the striking of the keys, and of every other electrical impulse or mechanical operation that occurs within the machine in the course of its carrying out the functions enshrined in its programming. All of these are, intrinsically, just meaningless physical events, and they get their significance as stages in the implementation of a program only because we take them to have such a significance.

But your computer's being a computer at all just consists of its implementing various programs; and its implementing such Programs

just consists of our taking it to be doing so, of our using it to run the programs. In itself, the machine is nothing more than a hunk of plastic, steel, silicon, and wires, with electrical current running through it. It counts as a computer, Searle suggests, only relative to us and our interests. Indeed, it is not strictly speaking a computer even then; it is we who literally compute when we use "computers." By the same token, it is we who really calculate when we use "calculators": the calculator itself is just a mechanical device, and the electrical current running through it, the images displayed on its screen) and the markings on its keypad are intrinsically without meaning. We give these things meaning and we do the calculating, with the device being merely an external aid, vastly different in degree of complexity from an abacus or a pencil and paper, but not (relevantly) different from them in kind.

For this reason, anything could in principle be used as a computer; all that matters is that the system thus used has a structure complex enough for us to be able to interpret its states as being stages in the program. To use an example of Searle's, the atomic structure of the wall of his study is complex enough for there to be some configuration of events taking place within it, at the micro-level, that could be interpreted as the implementation of a word processing program; in a sense, his wall is therefore "running" Word Perfect. Of course, we have no access to that system of micro-level events, so we could never actually find a workable way of isolating one part of the set of events and labeling it the "input," of isolating another part and labeling it the "output," and so on. But all that means is that we have no practical use for the wall as a potential word processor. Relative to our interests, it doesn't count as one, but in principle it could (and perhaps there might be creatures who would be able to make use of it). And the things that do count as word processors and the like do so only because we find it useful so to count them.

Computation, Searle concludes, is an observer-relative phenomenon. There is nothing intrinsic to the nature of anything in the material world that makes it a computer, or that makes it true that it is implementing a program. It is all a matter of interpretation: our interpretation. If we decide to count something as a computer, it is one; if not, then it isn't. There is nothing more to it than that. The most , complex machine that rolls off the assembly line at IBM will not count as a computer if we have no use at all for it; by contrast, even the pen sitting on the desk in front of you counts as a computer in the trivial sense that we can interpret it as "implementing" the following "program":'Lie there and don't move."

The problem Searle wants to pose for the computational conception of the mind should now be evident. If computation is observer-relative, then that means that its existence presupposes the existence of observers, and thus the existence of minds; so obviously it cannot be appealed to in order to explain observers or minds themselves. That would be to put the cart before the horse. It would be like trying to "explain"someone's appearance by appealing to a painting of her: "See, the painting looks like this; so that must be why she does too." Obviously, in this case, things are in reality the other way around: the painting's looking the way it does is to be explained in terms of the appearance of the person it is a painting of. By the same token, it is computation that must get explained in terms of the human mind, not the human mind in terms of computation. The brain is not intrinsically a digital computer, because nothing is. So the mind's ability to think in accordance with the laws of logic cannot be explained in terms of the brain's running a certain kind of program. The computational/representational theory of thought thus seems incoherent.

Another way to see the point is to recall that the computationalist account regards mental processes as the implementation of a set of algorithms.To implement an algorithm is to follow a set of explicit rules. As Hubert Dreyfus, another influential critic of computationalism, has pointed out, an apparent problem with the view that the mind can be explained entirely in terms of the following of some basic set of algorithmic rules is that any set of rules is capable of a variety of interpreiations. It is possible to fix the interpretation of a given set of rules by appealing to a set of higher-order rules, but that just pushes the problem back a stage, since these higher-order rules are themselves going to be susceptible to various interpretations. So, another way to understand Searle's argument is as follows: the fact that a computer is following some basic set of algorithmic rules cannot fully account for its behavior, because that the set of rules (and thereby its behavior) is to be understood in this way rather than that requires some interpretation to be put on those basic rules; and since there is, by definition, no more basic set to appeal to in order to fix the interpretation, we need to appeal to something outside the computer - a mind that interprets the rules. In that case, we cannot explain the mind itself in terms of the following of algorithmic rules, for that such rules are to be given this interpretation rather than that Presupposes the existence of a mind. Indeed, strictly speaking, that they truly count as rules at all presupposes that there is a mind interpreting them as rules; otherwise

all that is present are regularities of behavior that can be described as if they amounted to the following of rules.

Some have tried to reply to Searle's argument by noting that, strictly speaking, more is required of something if it is to count as a computer than merely that we could interpret some isolated set of its states as a computation. It is not enough, for example, for a system plausibly to count as implementing the computation "I + 2 = 3" that it has states that correspond to " l " and"2" which are followed by a state that corresponds to "3." For what it does genuinely to count as addition, it must also be true that had we instead counted the first two states as "3" and "4," the third state would have counted as "7" - and so on for other counter-factual inputs and outputs. But this does not seem to undermine Searle's basic point. All it shows is that a system is only going to be useful to us as a computer or calculator if it is complex enough to mirror all the possible computations we might want to perform with it, and not just some limited range. But this does not at all show that computation is not observer-relative. We couldn't make a knife out of just anything - steel and plastic will do, but shaving cream and butter won't - but that doesn't undermine the point that some- thing counts as a knife only relative to our interests. Not everything can effectively be used to express a word or sentence - ink marks and sounds will do, but cigarette smoke trails and water droplets are too formless and unstable - but that doesn't affect the point that a given physical object only counts as a word or sentence if we use it as a word or sentence. Similarly, a machine has to have a certain level of complexity if it is going to be useful to us as a word processor or calcuIator, but that doesn't change the fact that its being a word processor or calculator is ultimately a mind-dependent phenomenon.

These last examples indicate that if Searle is right, his argument would apply not only to the "computational" pait of the CRTT, but also to the "representational" part of it. The CRTT, as we've seen, holds that we think in a "language of thought,"where this language is realized in "sentences" somehow instantiated in the neural wiring of the brain. But as we've seen, physical shapes, patterns of sound, electrical impulses, and the like by themselves have no meaning. And the point is not merely that the word "cat" does not refer to cats apart from our taking it so to refer; it doesn't even count as a word in the first place, whatever we take it to refer to, unless are so count it. But the same is true of sentences. Nothing is intrinsically a sentence; something's status as a sentence is entirely relative to our using it as one. In itself, a sentence is just a string of marks on paper, a series of noises, or

whatever. And this seems no less true of neural wiring patterns: as one set of physical phenomena among others, they appear to have no intrinsic meaning or status as sentences, any more than do ink marks or sound-waves. But in that case, there cannot literally be sentences in our heads unless we interpret some neural processes occurring there as being instances of certain sentences - something which, quite obviously, happens only extremely rarely, if ever. More to the Point, if sentences too are observer-relative, then they cannot be appealed to in an explanation of the mind and its thoughts. If one accepts the basic thrust of Searle's position, then, the "representational" aspect of the CRTT seems as incoherent as the "computational" aspect.

Thought and consciousness

Finally, there is arguably a problem with the claim of the "language of thought,, hypothesis that the thoughts which have that language as their medium are never brought to consciousness - a claim that the theory must make, seeing as we are never aware of thinking in any such language, but only in the natural languages (English, German, French, Chinese, etc.) we use to speak. Searle argues that there can in principle be no such thing as an entity which is both literally a thought and totally unconscious. This is the "connection principle,"alluded to in the previous chapter, in which there is an inherent connection between something's being a thought and its being conscious. If this principle is true, it would seem to follow that there is yet another reason to regard the language of thought hypothesis, and the CRTT of which it forms a part, as incoherent.

Searle's argument for this principle brings into sharper focus the deep connections that, as we suggested in the previous chapter, seem to hold between consciousness, subjectivity, and intentionality. Boiled down to its essence, it goes like this: unconscious mental states, such as one's unconscious belief that water quenches thirst, have intentionality: in this case, the belief represents,is directed at,or is aboutthe fact of water's being thirst-quenching. But as with all intentional states, such unconscious states have "aspectual shape," in that they represent whatever it is they represent in some particular aspects rather than others. In the case at hand, the belief represents the fact in question as the fact that water is a thirst-quencher, and not necessarily as the fact that H_2O is a thirst-quencher (for the person who has the belief may know nothing about H_2O, and thus not know that water = H_2O). But aspectual shape is not something that can in principle be analyzed in

exclusively objective, third-person neurophysiological or behavioral terms. If we observe someone going to a spigot and turning it, there is nothing about this behavior by itself that determines conclusively that the person is seeking water rather than H_2O, for the behavior might be the same either way. Even asking him which one he is seeking won't be enough, because saying "I'm seeking water and not H_2O" won't by itself tell you whether what the person means by the sounds "water" and "H_2O" is the same as what you mean by those sounds. (And asking what the person does mean will just raise the same problem at another level: what does the person mean by these other sounds, which are made in order to explain what is meant by the first ones?)

The upshot, Searle concludes, is that it is only from the first-person point of view of the subjective experience of the person having the belief that the meaning of the person's words can be conclusively determined. It is important to note that Searle's claim isn't merely that we can't know for certain from the external, objective point of view what the meaning of the words is, but rather that there would be no fact of the matter at all what those words mean if the only evidence that existed was the external, third-person evidence alone. Here Searle appeals to a famous set of arguments given by the philosopher W. V. O. Quine (1908-2000) for what Quine called the indeterminacy of translation. Quine argued that an anthropologist who notes that a member of a previously unknown tribe constantly uses the expression "gavagai" in the presence of rabbits might naturally interpret that expression as meaning "rabbit," and go on to translate the rest of the speaker's language accordingly. But it is also possible, going by the speaker's behavior alone, that the expression could be translated instead as "undetached rabbit part" or "temporal stage of a rabbit" – assuming that the speaker's language reflects, unlike our own, a special interest in body parts that remain attached to the body, or in objects of ordinary experience considered as mere temporal stages of larger fourdimensional space-time structures (that is, the entire history of the rabbit from conception to death) - and that the rest of the speaker's language could be translated in light of these unusual assumptions. There is nothing in the speaker's behavior alone that could possibly favor one system of translation over the other, Quine argues, provided that each system of translation was thorough enough to account for all of the speaker's behavior. Quine, who was a kind of behaviorist – he held that there just is nothing to the mind over and above patterns of behavior - took this to have the startling consequence that there is no fact of the matter, period, about what any of us means whenever we

utter any expression: whether we decide to regard others, or even ourselves, as meaning "rabbit" or "temporal stage in the life of a rabbit" when we talk about rabbits, is entirely a pragmatic affair, a matter of which translation we find more useful. Neither interpretation is objectively closer to the truth than the other, for there is no objective truth of the matter in this case. Searle rejects this view utterly: there is, he insists, clearly more to the mind than behavior - there is also the subjective, first-person point of view of the conscious subject – and from this point of view a person does know that as a m6atter of fact it is, say, "rabbit" that he means, and not "temporal stage of a rabbit"'But Searle does agree with Quine that if third-person, behavioral (and neurophysiological) evidence were all we had to go on, there wouldn't be such a fact of the matter. The third-person, external evidence just isn't by itself enough to determine meaning - or, in particular, to determine aspectual shape.

If objective, third-person facts are not enough to determine aspectual shape, then they are also not enough to determine the content of an intentional mental state like a belief that water quenches thirst. But when a person has such a mental state unconsciously, such objective, third-person facts - facts about neural connections in the brain, about behavioral dispositions and the like - are all the relevant facts there are. So, strictly speaking, when he or she is not consciously aware of believing that water quenches thirst, he or she does not, in Searle's view, have that belief. But there is obviously a sense in which one has that belief even when one isn't conscious of it, isn't there? There is, Searle agrees, but what this amounts to is really just this: when someone isn't consciously entertaining that belief, what he or she has is a set of neural connections that have a tendency under certain circumstances to produce the conscious belief that water quenches thirst. Until the person is conscious of it, though, he or she doesn't literally have a mental state having the content that water quenches thirst; the person couldn't have it, given the inherent connection between the conscious, subjective, first-person point of view of the subject and the aspectual shape exhibited by all mental states involving intentionality.

If there is such an inherent connection there just couldn't be states which were literally mental and literally had intentionality, and yet were always in principle unconscious. That is, there couldn't be states of the sort the-"language of thought" hypothesis postulates: beliefs, desires, and so on, formulated in Mentalese. In Searle's view, if we are never conscious of such thoughts, we never really have them at all.

The defender of the CRTT could reply by suggesting that perhaps what we mean by "rabbit," and what we mean by anything else for that matter, really isn't as determinate from the first-person point of view as Searle thinks. Maybe you don't really know, even via introspection, precisely what you mean when you use "rabbit," or any other expression. And if not, there would be no reason to accept Searle's suggestion that an appeal to the subjective, first-person perspective of consciousness is necessary to account for the determinate meaning of our thoughts and expressions, for they just wouldn't have any determinate meaning in the first place.

This would, to say the least, seem to be a father extreme and counter-intuitive way to avoid Searle's conclusion - it appears to entail that there is no fact of the matter about whether you mean "rabbit" or "temporal stage of a rabbit': - and it brings us, at long last, to the issue of whether materialism can account for what seem to be the obvious facts about meaning or intentionality. The arguments considered in the previous chapter ler.. us to conclude that this is, ultimately, the key question the materialist has to face. The arguments of this chapter have reinforced this conclusion: the argument from reason implies that the standard materialist attempts to explain human rationality fail to account for the effect intentional mental states qua intentional have on the physical world; and Searle's various arguments suggest that the categories these materialist theories appeal to - computation, representation, language and its elements (for example, sentences) - presuppose intentionality and the point of view of the conscious subject, and thus cannot form the basis for a theory explaining the rational intentional processes of the subject. The last of his arguments has also reinforced the previous chapter's suggestion that there is an inherent link between consciousness, intentionality and subjectivity, and that one cannot account for one of these without accounting for the others. We will consider whether this argument is ultimately defensible as we focus on intentionality itself in the next chapter.

Further reading

An excellent introduction to many of the issues and arguments dealt with in this chapter is Tim Crane's The Mechanical Mind: A Philosophical Introduction to Minds, Machines, and Mental Representation, second edition (London: Routledge,2003). The claim that reasons are a species of causes is defended by Donald Davidson is his Essays on Actions and

Events (Oxford: Clarendon Press, 1980); the claim that they are not is defended by Wittgenstein's student G. E. M. Anscombe in her Intention (Oxford: Blackwell, 1959).

The language of thought hypothesis, and the computationa /representational theory of thought of which it is a part, are associated most famously with jerry Fodor, who has defended it in a series of publications. Particularly important are his The Language of Thought (Cambridge, MA: Harvard University Press, 1975) and Psychosemantics (Cambridge, MA: The MIT Press, 1987). Kim Sterelny's The Representational Theory of Mind: An Introduction (Oxford: Blackwell, 1990) is also helpful. Turing's ideas are presented in his famous essay "Computing Machinery and Intelligence," reprinted, with a number of other important articles relevant to the issues dealt with in this chapter, in Margaret A. Boden, ed. The Philosophy of Artificial Intelligence (New York: Oxford University Press, 1990).

Some important articles on the problem of mental causation are collected in John Heil andAlfred Mele, eds. Mental Causation (Oxford: Clarendon Press, 1995). The "argument from reason" has been presented in many different versions and by many different thinkers, most of whom did not call it by that name. C. S. Lewis is often cited as its inventor, though it seems that other people have independently developed similar ideas, both before Lewis and after. In any event, Lewis's version of the argument is to be found in his book Miracles (Macmillan, 1978), and is developed and defended by Victor Reppert in C. S. Lewis's Dangerous ldea: In Defense of the Argument from Reason (Downers Grove: InterVarsity Press, 2003). William Hasker's version is presented in chapter 3 of The Emergent Self (Ithaca: Cornell University Press, 1999). Karl Popper's related argument is in chapter 6 of Objective Knowledge, revised edition (Oxford: Clarendon Press, 1979). Alvin Plantinga's is presented in chapter 12 of Warrant and Proper Function (New York Oxford University Press, 1993) and debated in James Beilby, ed. Naturalism Defeated?: Essays on Plantinga's Evolutionary Argument against Naturalism (Ithaca: Cornell University Press, 2002).

Searle's Chinese room argument was originally presented in "Minds, Brains, and Programs," which has been very widely reprinted (including in the Boden anthology cited above). That article and a number of early responses can be found together in Rosenthal's anthology The Nature of Mind, referred to in earlier chapters. Searle's ideas on the observer-relativity of computation and the "connection principle" are developed most thoroughly in The Rediscovery of the Mind (Cambridge, MA: The MIT Press, 1992). All of these ideas are

debated in John Preston and Mark Bishop, eds.views into the Chinese Room: New Essays on Searle. and Artificial Intelligence (Oxford: Clarendon Press, 2002). Dreyfus's views are developed most thoroughly in what computers still can't Do (cambridge, MA: The MIT Press, 1992). Quine's argument is most thoroughly developed in his Word and Object (Cambridge, MA: The MIT Press, 1960).

Another important challenge to the computationalist model of the mind is, in the view of some writers, posed by Godel's famous incompleteness results in mathematical logic. An argument to that effect was first proposed by J. R. Lucas in his "Minds, Machincs, and Godel'" available in Alan R. Anderson, ed" Minds and Machines (Englewood Cliffs, NJ: Prentice Hall, 1964) and developed at length by Roger Penrose in The Emperor's New Mind (New York: Oxford University Press,1989).