15

Minds and Brains Without Programs

John Searle

John Searle was born in Denver in 1932. He has been Professor of Philosophy at the University of California, Berkeley since 1959. Before that he taught at Oxford where he studied as an undergraduate and graduate student. He is best known for his contribution to linguistic philosophy, in particular his work on the theory of speech acts. He is married with two sons and lives in the foothills of Berkeley with a collection of antique rugs, several cars, much fine wine from the vineyard he helps to run, and a small dog called Russell.

The Gap

The aim of this chapter is to present an interim report on some arguments I have been having with philosophers and people in various other disciplines.[1] I want to begin by placing the issues in a somewhat larger context.

There is a remarkable lacuna in twentieth-century intellectual life. ('Lacuna' is perhaps a euphemism for 'scandal'.) We are quite confident that we can give explanations of human behaviour in ordinary, common-sense terms. So we say such things as, 'That man voted for Ronald Reagan because he thought Reagan would cure inflation.' Such remarks are part of common-sense or grandmother psychology. To give it a fancy name, we could call it 'intentionalistic psychology'. We also suppose that underlying this level of explanation there must be a neurophysiological level of explanation. But we really don't know how to give neurophysiological explanations of ordinary human behaviour. We don't know how to make such claims as, 'The man voted for Reagan because of a condition in his thalamus.' This leaves us in an intellectually embarrassing situation. We are reasonably confident in using grandmother psychology at the higher level and we think there must be a hard science underlying it at the lower level, but we haven't the faintest idea how the lower level works in explaining specific cases of normal human behaviour. We use grandmother psychology all of the time, but we are embarrassed to call it a science. Nobody, for example, has the nerve to go to the National Science Foundation and ask for a grant to do grandmother psychology. Yet we don't know enough about the lower level to make it work. So it seems we have a gap.

$$\frac{\text{intentionalistic psychology}}{\text{neurophysiology}} \leftarrow \text{gap}$$

A Chinese Room (British Library Department of Oriental Manuscripts 15530.c.11).

Some of the great intellectual efforts of the twentieth century have been attempts to get the gap filled – to find something that would be a science of human behaviour, but wasn't common-sense psychology and wasn't neurophysiology either. And if you live long enough, it is interesting to look back and see all the dead carcasses of theories that were supposed to fill the alleged gap. In my lifetime the most spectacular failure was behaviourism. But I also lived through several other failed efforts. There was games theory and there was information theory. I don't suppose anyone reading this is old enough to remember cybernetics, but at one time great claims were made for the future of cybernetics. There was something called 'structuralism', and that was followed by something called 'post-structuralism'. And now there is sociobiology, yet another candidate to fill the gap.

However, the leading candidate right now is called 'cognitive science', and the central research programme in cognitive science is often thought to be artificial intelligence. There are different schools of cognitive science and artificial intelligence, but the most ambitious gap-filling theory is the one that says that work in cognitive psychology and artificial intelligence has now established that the mind is to the brain as the computer program is to the computer hardware. This is a very common equation in the literature: mind/brain = program/hardware. To distinguish this view from more cautious versions of artificial intelligence, I have labelled it 'strong artificial intelligence' ('strong AI' for short). According to strong AI, the appropriately programmed computer with the right inputs and outputs literally has a mind in exactly the same sense that you and I do.

Now this view has something interesting consequences. It has the consequence, for example, that there is nothing essentially biological about the human mind. It so happens that the programs which are constitutive of minds are run in the wetware that we have in our biological machine, our biological computer in the head. But those very same programs could be run on any hardware computer whatever that was capable of sustaining the program. And that has the further consequence that anything whatever, any system whatever, could have thoughts and feelings – and indeed it not only *could have*, but *must have*, thoughts and feelings – in exactly the same sense that we do, provided only that it is running the right program. That is, if you have the right program with the right inputs and the right outputs, then any system running that program, regardless of its chemical structure (whether it is made out of old beer cans or silicon chips or any other substance) *must have* thoughts and feelings in exactly the same way you and I do. And this is because that is all there is to having a mind: having the right program. Now, whenever I attack this view, many people say, 'But surely nobody can believe that.' I'm going to tell you the names of some of the people who believe this, so you won't think I am just attacking a straw man.

Herbert Simon of Carnegie-Mellon University has written on a number of occasions that we already have machines that can literally think, that can think in the same sense that you and I do. Philosophers have been worried for centuries about whether or not you could build a machine that could think, and now we learn that they do it every day at

Carnegie-Mellon. Simon's colleague Alan Newell, in a lecture I heard him give in San Diego at the founding meeting of the Cognitive Science Society, said we have discovered (it is not just some hypothesis we are considering, but we have 'discovered') that intelligence is purely a matter of physical symbol manipulation. So any machine that is capable of manipulating the right symbols in the right way literally has intelligent processes in exactly the same sense that you and I do. Marvin Minsky says that the next generation of computers will be so intelligent that we will be lucky if they keep us around the house as household pets. And Freeman Dyson is quoted in the *New York Times* as having said that since we now know that mental processes, such as consciousness, are purely formal processes, there is an evolutionary advantage to having these formal processes (consciousness, and so on) go on in silicon chips and wires, because that kind of stuff is better able to survive in a universe that is cooling off than organisms like us made out of our messy biological machinery. So the next stage in evolution, on this view, will be made out of wires and silicon. My all-time favourite in this literature (and I recommend the literature to you because it is marvellous) is from John McCarthy, the inventor of the term 'artificial intelligence'. McCarthy has written: 'Machines as simple as thermostats can be said to have beliefs . . .' And indeed, he adds, 'having beliefs seems to be a characteristic of most machines capable of problem solving performance'.[2] So I asked him, 'John, what beliefs does your thermostat have?' I admire his courage. He said, 'My thermostat has three beliefs. My thermostat believes it's too hot in here, it's too cold in here, and it's just right in here.'

Now, I like this thesis for a very simple reason. This equation, mind/brain = program/hardware, is unusual in philosophy because it is a reasonably clear thesis. You can state it with reasonable precision. And unlike most philosophical theses, it is subject to a very simple and, I think, decisive refutation. This is a refutation that I have published elsewhere, but I am going to repeat it briefly because it is not universally accepted in the artificial intelligence community that I have in fact refuted this view. Then I want to go into a deeper issue, and that is this: one of the reasons people believe in strong AI is that they can't see any other way to solve the mind–body problem. I am convinced that one of the sources of the belief that all there is to having a mind is having a computer program is that people can't see any other way of solving the mind–body problem without resorting to dualism. The challenge is often presented to me: 'Well, if you don't accept the artificial intelligence analysis of the mind, then what is your solution to the mind–body problem? Aren't you forced into dualism or mysticism or vitalism or some equally weird view?' So really I have two tasks. I want to refute strong artificial intelligence, and I want to solve the mind–body problem.

The Chinese Room Revisited

The argument against strong AI, I fear, is rather simple. This argument occurred to me when I read Schank and Abelson's book about their story-understanding programs.[3] Some of you will be familiar with this, but I will

go over the steps of how their programs work. These very ingenious programs have been designed at Yale University. The programs do what they call 'understanding stories'. The computer is given a very simple story as input. A typical story would be something like the following:

'A man went into a restaurant and ordered a hamburger. When they brought him the hamburger it was burned to a crisp. The man stormed out of the restaurant without paying for the hamburger.'

Then, you ask the computer, 'Did the man eat the hamburger?' And lo and behold, the computer says as output, 'No, the man did not eat the hamburger.' Or you give the computer another story:

'A man went into a restaurant and ordered a hamburger. When he was served the hamburger he was delighted with it, and when he left the restaurant he paid the bill and left a large tip for the waitress.'

If you then ask the computer, 'Did the man eat the hamburger?', the computer says, 'Yes, the man ate the hamburger.' Now notice, in neither story did it say explicitly whether or not the man ate the hamburger. How does it work? It works because the program has in its data base what is called a 'restaurant script'. The restaurant script is a representation of how things normally go on in restaurants. When the computer gets a story, it matches the story against the restaurant script, and then when it gets the question about the story, it matches the question against both the story and the restaurant script. Since it 'knows' how things are supposed to go on in restaurants, it can produce the right answer. The claim that is often made about the programs is that since the machine satisfies the

Figure 15.1 Alan Turing (1912–54) as a young man. He was a pioneer in computer theory whose ideas anticipated the development of artificial intelligence. He conceived of a test (the Turing test) to discover whether a computer has achieved human-level intelligence (photograph Andrew Hodges).

Turing test, the machine must literally understand the story.[4] It must literally understand the story in exactly the same sense that you and I would understand such stories if we were asked such questions and gave good answers.

The Turing Test

The Turing test was designed by Alan Turing to test whether a computer or other system had the same cognitive abilities as humans. As it is usually understood, the test is whether or not an expert would be able to distinguish the machine's performance from that of a human. If not, the machine has the same cognitive abilities as the human. See Roger Penrose's account on p. 261.

It seems to me that there is a very simple refutation of this claim. The refutation is just to imagine that you are the machine. I like to imagine it the following way.

Suppose I am locked in a room. In this room there are two big bushel baskets full of Chinese symbols, together with a rule book in English for matching Chinese symbols from one basket against Chinese symbols from the other basket. The rules say things such as, 'Reach into basket 1 and take out a squiggle-squiggle sign, and go put that over next to the squoggle-squoggle sign that you take from basket 2.' Just to look ahead a moment, this is called a 'computational rule defined over purely formal elements'. Now let us suppose that the people outside the room send in more Chinese symbols together with more rules for shuffling and matching the symbols. But this time they also give me rules for passing back Chinese symbols to them. So, there I am in my Chinese room, shuffling these symbols around; symbols are coming in, and I am passing symbols out according to the rule book. Now, unknown to me, the people who are organizing all of this on the outside of the room call the first basket 'a restaurant script', the second basket 'a story about the restaurant'; the third batch of symbols they call 'questions about the story', and the symbols I give back to them they call 'answers to the questions'. The rule book they call 'the program', themselves they call 'the programmers', and me they call 'the computer'. Now after a while, suppose I get so good at answering these questions in Chinese that my answers are indistinguishable from those of a native Chinese speaker. All the same, there is an important point that needs to be emphasized. I don't understand a word of Chinese, and there is no way that I could come to understand Chinese from instantiating a computer program in the way that I described it. And this is the point of the story: *if I don't understand Chinese in that situation, then neither does any other digital computer solely in virtue of being an appropriately programmed computer, because no digital computer solely in virtue of its being a digital computer has anything that I don't have.* All that a digital computer has, by definition, is the instantiation of a formal computer program. But since I am instantiating the program, since

we are supposing we have the right program with the right inputs and outputs, and I don't understand any Chinese, then there is no way any other digital computer *solely in virtue of instantiating the program* could understand Chinese.

Now that is the heart of the argument. But the point of the argument, I think, has been lost in a lot of the subsequent literature developed around this, so I want to emphasize the point of it. The point of the argument is not that somehow or other we have an 'intuition' that I don't understand Chinese, that I find myself *inclined to say* that I don't understand Chinese but, who knows, perhaps I really do. That is not the point. The point of the story is to remind us of a conceptual truth that we knew all along; namely, that there is a distinction between manipulating the syntactical elements of languages and actually understanding the language at a semantic level. What is lost in the AI *simulation of* cognitive behaviour is the distinction between syntax and semantics.

Now the point of the story can be stated more generally. A computer program, by definition, has to be defined purely syntactically. It is defined in terms of certain formal operations performed by the machine.[5] That is what makes the digital computer such a powerful instrument. One and the same hardware system can instantiate an indefinite number of different computer programs, and one and the same program can be run on different hardwares, because the program has to be defined purely formally. But for that reason the formal simulation of language understanding will never by itself be the same as duplication. Why? Because in the case of actually understanding a language, we have something more than a formal or syntactical level. We have a semantics. We do not just shuffle uninterpreted formal symbols, we actually know what they mean.

You can see this by enriching the argument slightly. There I am in the Chinese room shuffling these Chinese symbols. Now suppose that sometimes the programmers give me stories in English and ask me questions, also in English, about these stories. What is the difference between the two cases? Both in the English case and in the Chinese case, I satisfy the Turing test. That is to say, I give answers which are indistinguishable from the answers that would be given by a native speaker. In the case of Chinese, I do that because the programmers are good at designing the program, and in the case of English, I do that because I am a native English speaker. What is the difference, then, if my performance is equivalent in the two cases? It seems to me that the answer to that question is obvious. The difference is that I know English. I know what the words mean. In the case of English, I don't just have a syntax, I have a semantics. I attach a semantic content, or meaning, to each of these words; and therefore I am doing something more than a digital computer can do just in virtue of instantiating a program. I have an interpretation of the words, and not just the formal symbols. Notice that if we try to give the computer an interpretation of the formal symbols, all we can do is give more formal symbols. All we can do is put in more uninterpreted formal symbols. By definition, the program is syntactical, and the syntax by itself is never sufficient for the semantics.[6]

Well, that is my rejection of this equation, mind/brain = program/hardware. Instantiating the right program is never sufficient for having a

mind. There is something more to having a mind than just instantiating a computer program. And the reason is obvious. Minds have mental contents. They have semantic contents as well as just a syntactical level of description.

There is a persistent misunderstanding of my argument which I wish to block immediately. Some people suppose that I am claiming that is is in principle impossible for silicon chips to duplicate the causal powers of the brain. That is not my argument; indeed, it has no connection whatever with my argument. It is a factual question, not to be settled on purely philosophical or a priori grounds, whether or not the causal powers of neurons can be duplicated in some other material, such as silicon chips, vacuum tubes, transistors, beer cans, or some quite unknown chemical substances. The point of my argument is that you cannot duplicate the causal powers of the brain solely in virtue of instantiating a computer program, because the computer program has to be defined purely formally. It is important to emphasize that artificial intelligence, whether strong or otherwise, has nothing whatever to do with the chemical properties of silicon or any other substance. Once the AI partisan concedes that these are even relevant, he has abandoned the thesis of AI. AI is about the 'cognitive' powers of programs. It has nothing whatever to do with the specific chemical properties of hardware realizations of programs.

However, that leaves us with the second question. If we reject the equation and we reject AI as a gap-filler, then what is our analysis of the relationship between the level of intentionality and the level of neurophysiology? Well, the short answer is that the reason the gap-fillers always fail is that there isn't any gap to fill. There isn't any gap between the level of intentionalistic explanations and the level of neurophysiological explanations. But in order to substantiate that, I need, as I promised earlier, to solve the mind–body problem.

Four Puzzles

Before confronting the 'mind–brain problem' head-on I want to step back a minute and ask why this problem has seemed so intractable. Why in philosophy, psychology and neurophysiology, do we still have a mind–body problem? Since Descartes, at least, the general form of the mind–body problem has been the problem of accommodating our common-sense and pre-scientific beliefs about the mind to our general scientific conception of reality. Our scientific conception of the world as a physical system or as a set of interacting physical systems has grown in power and comprehensiveness, and it has seemed increasingly difficult to find any place for mind in this conception. Some of the pre-scientific views that appear to be challenged by the growth of a scientific world-view derive from religion or morality – doctrines such as the immortality of the soul, the freedom of the will, the nature of moral responsibility – and about these issues I will have nothing to say in this discussion. I will be concerned with the narrower, and I believe more pressing, question, how can we square what we know, or seem to know, about the world in general with what we know, or seem to know, about the operation of our own

minds? Quite apart from the speculations of religion and the presuppositions of morality, we know a number of things about our minds, and my aim is to give a coherent account of the relationships between what we know about our own minds and what we know about the way the world works in general. Why then has this narrower, non-religious, non-moral problem seemed so intractable? Why, to repeat, is there still a mind–brain or mind-body problem?

The features of our common-sense conception of the mind that seem hard to assimilate to our general scientific conception of the world are at least the following four:

Consciousness

I, at the moment of writing this, and you, at the moment of understanding it, are both conscious. It is just a plain fact about the world that it contains conscious mental states, but it is hard to see how (mere) physical systems could have consciousness. How could such a thing occur? How, for example, could this hunk of grey and white matter inside my skull be conscious?

Intentionality

Many of my mental states, such as, for example, my beliefs and desires and my visual perceptions and my intentions, are directed at, or about, or of objects and states in the world apart from themselves. This feature, called 'intentionality', is characteristic of human minds. But again, how could such a thing occur? How could processes in my brain, which after all consists, in the end, of 'atoms in the void', be *about* anything? How can atoms in the void *represent* anything? One is inclined to say: things and processes in the world just are; whether we are thinking of biological *processes* such as digestion and sequences of neuron firings or ordinary physical *things* such as stones and trees, it seems quite impossible that any of these should be *about* anything. How can *aboutness* be an intrinsic feature of the world?

Subjectivity

Mental states are characteristically subjective. But it is hard to understand how the objective physical world, equally open to all competent observers, should contain anything essentially subjective such as, for example, conscious mental states. Naively construed, the subjectivity of mental states is marked by such facts as that I have my states and not yours; mine are accessible to me in a way they are not accessible to you; I perceive the world from my point of view, not from your point of view, etc. How can subjectivity be a real part of the world?

Intentional Causation

Even if there were such things as mental states, it is hard to see how they could make any real difference in the world. Could anything, so to speak, as 'gaseous' and 'ethereal' as a conscious mental state have any impact on

a physical object such as a human body? How could mental phenomena ever push objects around or have any other physical significance? Wouldn't mental states, even if they existed, be just epiphenomenal?

Let us call these problems, respectively, the problems of consciousness, intentionality, subjectivity and intentional causation. Though not all mental states have all of these four features, they are none the less real and typical features of mental phenomena. We know, for example, that people are often in a state of consciousness, that, for example, they often have thoughts and feelings which refer to objects and states of affairs outside themselves, that they apprehend the world from a subjective point of view, and that their thoughts and feelings make a difference to their behaviour. I believe any account of the mind–brain problem must, at a minimum, be able to account for all of these facts.

On the view of mental states adopted in this essay, mental states and processes are real biological phenomena in the world; as real as digestion, photosynthesis, lactation or the secretion of bile. The aim of this chapter is not to show in detail how such biological phenomena are related to the neurophysiological processes of the brain – no one knows in detail how they are related – rather its aim is the more modest one of showing how it is even possible that mental states could be biological phenomena in the brain. I believe it is a typical but unstated tacit assumption behind many of the implausible contemporary doctrines concerning the mind – doctrines such as behaviourism or strong artificial intelligence – that it is simply impossible to accommodate a naive common-sense account of the mind with an overall scientific world-view. And I believe that it is the sense of desperation caused by the feeling that no coherent account can be given which accommodates common-sense mentalism with hard science that leads people to say the implausible, and sometimes silly, things they say about the nature of the mind. The view that I am about to expound of the relation of mind and brain is consistent with what is known about brain functioning and also consistent with a general biological approach to biological phenomena. My approach does not, like strong AI, try to treat the mind as something formal or abstract; nor does it, like certain forms of functionalism, try to treat the mind as simply a neutral set of causal powers with no intrinsically mental characteristics. Frankly, I think that the approach I am about to present is pretty much an obvious and common-sense view, and until I got involved in these recent polemics, I had assumed that it was widely accepted, so widely accepted as to hardly be worth a separate statement. Nonetheless, my previous formulations of it have been characterized by my critics as 'mysticism' (Ringle)[7], 'sophistry' (Dennett)[8], 'religious' (Hofstadter)[9], etc. Perhaps, therefore, it is worth spelling out the position in some detail so that anyone can see that these charges are quite unfounded. I need hardly emphasize that I am not the first person to hold this sort of view, and similar biological approaches to the mind–body problem can be found at least as far back as the nineteenth century.

The Brain and its Mind

How does the brain work? In detail, no one knows. I have an amateur's ignorance of the subject, but even the best experts are up to the present

time baffled by what one would think are the most fundamental questions. What exactly is the neurophysiology of consciousness? Why do we need sleep? How exactly are memories stored in the brain? Why does alcohol make us drunk? Why does aspirin relieve pain? As recently as 1978, a famous neurophysiologist, David Hubel, wrote: 'There are [areas of the brain] the size of one's fist, of which it can almost be said that we are in the same state of knowledge as we were with regard to the heart before we realized that it pumped blood.'[10] Furthermore, in our ignorance, we grope for metaphor and analogy, usually based on the latest technology. Thus, nowadays, the most fashionable view is that the brain is a digital computer, but in my childhood I was assured that it was a kind of telephone switchboard; Charles Sherrington compared the brain to a telegraph system and to a jacquard loom; Sigmund Freud compared it to hydraulic pumps and electromagnetic systems; Leibniz compared it to a mill and I am told that certain Ancient Greeks thought the brain functioned like a catapult. The very latest view among neurophysiologists is that the brain functions like a Darwinian natural selection system.

However, though there is much to learn, we are not totally ignorant, and in a discussion such as this we need to remind ourselves of a few elementary things about the brain. Like all organs, the brain consists of cells. However, unlike other organs, the brain and the rest of the nervous system consist in large part of very special kinds of cells, neurons. By current estimates there are probably between 50 and 100 billion neurons in the human brain. Neurons come in a bewildering variety of types, but the typical garden-variety neuron consists of a cell body, or soma, with two types of long fibres sticking out of it, a single axon, and a number of dendrites. Neurons come in contact with each other at certain small bumps called synapses. The axons and dendrites don't actually fuse together at synapses but the axon characteristically has on it a little protuberance, the bouton, that abuts on to the dendrite, the tiny gap between them being the synaptic cleft. There are also synapses on the soma. Some neurons in the cerebellum have as many as 200,000 synapses on one cell. One of the basic functions of the neuron is the transmission of electrical impulses, that is, brief, 'all-or-nothing' changes in electrical potential. Each electrical impulse passes from the soma along the axon. However, in most neurons the electrical impulse does not pass directly from one neuron to the next; rather the electrical impulse, upon reaching the bouton, causes the release of small amounts of fluid from little compartments in the bouton, the synaptic vesicles, into the synaptic cleft. The release of these fluids, the neurotransmitters, at the synapses, can have either an excitatory or an inhibitory effect on the next neuron in line. If excitatory, it will tend to cause the next neuron to fire or increase its rate of firing. If inhibitory, it will tend to prevent the neuron from firing or decrease its rate of firing. From a functional point of view, the important thing is not that the neuron fires, because many neurons fire all the time anyway. What is important are the variations in the *rate* of neuron firings; specifically, variations in the rate of axon firing from the sum of excitations and inhibitions in dendrites.

Figure 15.2 A typical garden variety neuron. Otto Deiters drew this picture, remarkably accurate for the time, in 1865. He carefully dissected individual nerve cells from the spinal cord of an animal (probably an ox). The drawing shows the cell body (with its nucleus inside), the long axon and the numerous branching dendrites. Deiters thought that he saw other fine axons sprouting from the dendrites, but these were probably the terminations of the fibres of other neurons ending in synapses on the dendrites.

It is important to emphasize this point because several authors have erroneously supposed that the all-or-nothing character of the firing of

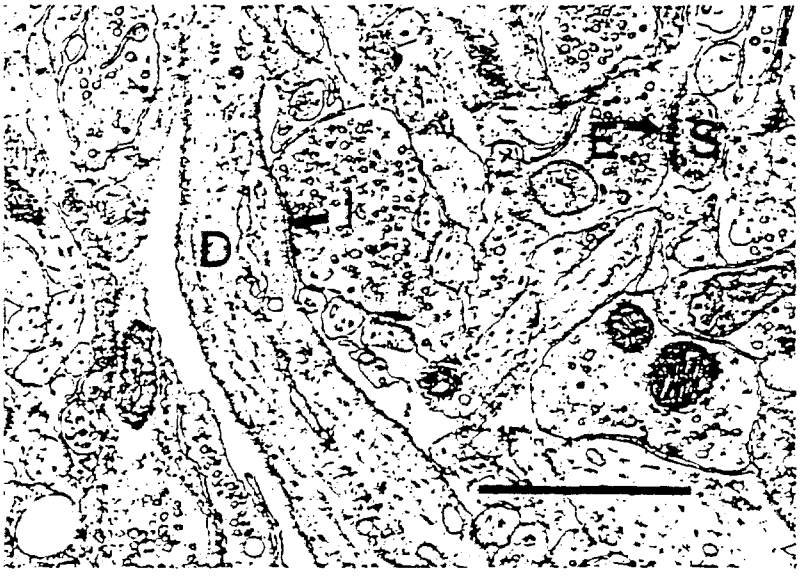

Figure 15.3 Connections between neurons in the cerebral cortex of a rat. This electron micrograph (prepared by Alan Larkman) shows an enormously magnified view of the tangle of axons and dendrites in the cortex. The horizontal bar at the bottom right represents one thousanth of a millimetre. The shape labelled D is one of the dendrites (see fig. 15.2) of a cell in the cortex. It is contacted by the terminal of an axon, marked I, which contains many tiny oval vesicles containing a chemical transmitter substance, which is released into the narrow gap between terminal and dendrite, in the region shown by the arrow, when each nerve impulse arrives at the terminal. The membranes of the terminal and the dendrite are equally thickened in the area of contact, which is thought to be characteristic of an inhibitory synapse, whose transmitter substance decreases the activity in the next nerve cell. Nearby is a clear view of another synaptic contact between an axon terminal (marked E) and a round knob-shaped process – probably one of the 'spines' which cover the dendrites of many cortical cells. In this case the membrane of the axon is much less thickened than that of the spine in the region of contact (arrow), which indicates that this is an excitatory synapse.

nerve impulses constitutes evidence that the principles of brain functioning are those of a digital computer.[11] Nothing could be further from the truth. As far as we know, the functional aspect of the neuron is the non-digital variation in the rate of firing.

On the traditional account of the brain, the account that takes the neuron as the fundamental unit of brain functioning, the remarkable thing about the relationship between the brain and the mind is simply this. All of the enormous variety of inputs that the brain receives – the photons that strike the retina, the sound waves that stimulate the sensory cells of the inner ear, the pressure on the skin that activates nerve endings for pressure, heat, cold and pain, etc. – all of these inputs are converted into one common medium: variable rates of neuron firings. Furthermore, and equally remarkable, these variable rates of neuron firing relative to different neuronal circuits and different local conditions in the brain produce all of the variety and heterogeneity of the mental life of the human or animal agent. The smell of a rose, the experience of the blue of the sky, the taste of onions, the thought of a mathematical formula – all of these are produced by variable rates of neuron firing, in different circuits

relative to different local conditions in the brain. Now what exactly are these different neuronal circuits and what are the different local environments that account for the differences in our mental life? In detail, no one knows, but we do have good evidence that certain regions of the brain are specialized for certain kinds of experiences. The visual cortex plays a special role in visual experiences, the auditory cortex in auditory experiences, etc. Vision is one of the best understood (or least inadequately understood) functions of the brain and in the case of vision there appear to be quite specialized neurons in the visual cortex capable of responding to specific different features of visual stimuli. Suppose that auditory stimuli were fed to the visual cortex and visual stimuli were fed to the auditory cortex. What would happen? As far as I know, no one has ever done the experiment, but it seems reasonable to suppose that the auditory stimulus would be 'seen', that is, that it would produce visual experiences, and the visual stimulus would be 'heard', that is, it would produce auditory experiences, in both cases because of specific, though largely unknown, features of the visual and auditory cortex respectively. Though this hypothesis is speculative, it has some independent support if you reflect on the fact that a punch in the eye produces a visual flash ('seeing stars') even though it is not an optical stimulus.

In my layman's view the amount of knowledge we now have about the nature and functioning of neurons is quite impressive: however, there is now a lot of evidence that, in order to understand the role of the brain in mental life, we need to understand the functioning of the brain at higher levels than that of individual neurons, and in particular among the various higher levels we need to understand the functioning of systems of neurons organized into neural networks or neural circuits. For many functions of the brain the unit of functioning is not the single cell but the network of cells and, at this level, brain functioning is a matter of interaction between a large set of neural networks. The best anatomical evidence for the existence of more or less independently functioning networks, at least in the cortex, is the existence of neural modules in the form of vertically oriented columns, cylinders or slabs of cells in the cortex.[12] As Gerald Edelman says, 'The greatest achievement in thinking about the cortex, the greatest revolution, is that the cortex is not just a continuous horizontally disposed sheet, but is vertically organized as stacks of slabs or columns.'[13] These modules may vary in the number of neurons they contain, from as few as 50 neurons to as many as 50,000 or even more. On the modular view the real significance of the neuron is in the contribution it makes to the functioning of the module.

I have no idea if the combination of the neuronal account and the modular account or perhaps some other account is the correct account of brain functioning. But one conclusion emerges clearly from even the most cursory investigation of the functioning of the brain: *mental phenomena, whether conscious or unconscious, whether visual or auditory, pains, tickles, itches, thoughts, and all the rest of our mental life, are caused by processes going on in the brain.* Mental phenomena are as much a result of electrochemical processes in the brain as digestion is a result of chemical processes going on in the stomach and the rest of the digestive tract. I think this is an obvious fact about how the world works and yet its full implications are

not generally realized by students of artificial intelligence, cognitive science or philosophy. It is also important to emphasize that the relevant causal processes are entirely internal to the brain. Though *in fact* mental events mediate between external stimuli and motor responses, there is no *essential* connection. A man could, for example, have a terrible pain without having either a pain stimulus of the peripheral nerves or any pain behaviour. This simple fact is sufficient to discredit the entire behaviourist tradition in philosophy.

To substantiate this at least a little bit, let us tell a part of the causal story for one type of conscious mental phenomenon: pain. Pain signals are transmitted from sensory nerve endings to the spinal cord by two types of fibres: A delta fibres which are specialized for prickling sensations and C fibres which are specialized for burning and aching sensations. In the spinal cord they pass through the tract of Lissauer and terminate on the neurons of the cord. As the signals go up the spine and enter the brain

Figure 15.4 René Descartes (1596–1650) conceived of the body, including the brain, as a machine, capable of automatic (reflex) reactions to events in the world. This illustration, from the 1662 edition of the Treatise of Man, *shows the way in which the heat of a fire (A) causes disturbance of sensory nerve fibres in the hand (B), and transmission of a message up to the brain, 'just as, pulling on one end of a cord, one simultaneously rings a bell which hangs at the opposite end'. Descartes thought that this tugging would in turn open pores in the walls of the fluid-filled cavities (the ventricles) of the brain, causing 'animal spirit' to flow out through what Descartes believed were hollow tubes in the centres of other nerves connected to the muscles of the body. This liquid was then thought to inflate the muscles and cause various reflex reactions such as withdrawing the injured hand or moving the other arm across to protect it. This description, even if adequate to explain the* behaviour *associated with pain, says nothing about the* sensation *itself. That was supposed to be caused by simultaneous disturbance of the pineal gland (the oval shape at the back of the head), which acted as the interface between the material body and the spiritual soul.*

they separate into two pathways, the prickling pain pathway and the burning pain pathway. Both pathways pass through a structure called the thalamus but, beyond that level, prickling pain is more clearly localized in the somatic sensory cortex, specifically, somatic area 1, whereas the burning pain pathway transmits signals not only upwards into the cortex but also laterally into the hypothalamus and other basal regions of the brain. In consequence of these differences it is much easier to localize a prickling sensation, whereas burning and aching pains can be more distressing, perhaps because they activate more of the nervous system. The actual sensation of pain appears to be caused both by the stimulation of the basal regions, especially the thalamus and the stimulation of the somatic sensory cortex.

Now for philosophical purposes, it is essential to hammer this point home: sensations of pain are caused by a series of events that begin at free nerve endings and end in the thalamus and other regions of the brain. Indeed, as far as the actual sensations are concerned, the events inside the central nervous system are quite sufficient to cause pains, as we know from both phantom limb pains,[14] and from the pains caused by artificially stimulating relevant portions of the brain. And what is true of pain is true of mental phenomena generally. To put it crudely, and counting the rest of the central nervous system as part of the brain for the purposes of this discussion, everything that matters for our mental life, all our thoughts and feelings are caused by processes inside the brain. As far as the causation of mental states is concerned, the crucial step is the one that goes on inside the head, and not the external stimulus. And the argument for this is simply that if the events outside the brain occurred but caused nothing in the brain, there would be no mental events, whereas if the events in the brain occurred the mental events would occur even if there were not outside stimulus.

I believe these points are obvious, but they are inconsistent with two very common views about the mind. One view treats external causation as the essential form of causation for mental contents. But on the present account the external causal chains are only important to the extent that they actually impact on the central nervous system. Another widely held view is that there cannot be a causal relation between mind and brain states, because mental states just are brain states and the form of the identity relation in question precludes the possibility of any psychophysical causal relations. For example, many materialist philosophers used to claim that pains just *are* C fibre stimulations, whereas on the present view, C fibre stimulations are not *identical* with pains but are *part of the causes* of (certain kinds of) pain.

But then let us ask the next obvious question: if pains and other mental phenomena are caused by neurophysiological processes, what are the phenomena themselves? Well, in the case of pains, they are obviously very unpleasant sorts of sensations; but that answer leaves us unsatisfied because it doesn't, so to speak, tell us how to locate pains and other mental phenomena relative to the rest of the world we live in. How do pains fit into our overall ontology? Once again, I think the answer to this question is obvious, though it will take some spelling out. To our first claim, namely that pains and other mental phenomena are caused by brain

processes, we need to add a second claim: *pains and other mental phenomena are features of the brain.*

One of the primary aims of this discussion is to show how *both* of these propositions can be true together. There are different levels of philosophical puzzlement that such a pair of theses can generate. At one level we can be puzzled as to how mental and physical phenomena can stand in causal relations when one is a feature of the other. Wouldn't it lead to the dreaded doctrine of *causa sui*? That is, wouldn't it imply that the mind caused itself? At bottom most of our puzzlement comes from a misunderstanding of the nature of causation. It is tempting to think that whenever A causes B there must be two discrete events, one identified as the cause, the other identified as the effect; that all causation functions on the model of a bolt of lightning causing a clap of thunder; and if we have this crude model of causation we will be tempted to think that the causal relations between the brain and the mind force us to accept some kind of dualism; that events in one realm, the 'physical', cause events in another realm, the 'mental'. But this seems to me a mistake. And the way to remove the mistake is to get a more sophisticated concept of causation. To do this, let us turn our attention away from mind–brain relations for a while to observe some other sorts of causal relationships in nature.

Macro- and Micro-properties

A common distinction in physics is between micro- and macro-properties of systems. Consider, for example, the desk in front of me, or the creek that flows outside my office window. Each system is composed of micro-particles and the micro-particles have features at the level of molecules and atoms as well as subatomic particles. But each system also has certain properties such as solidity in the case of the table, or liquidity in the case of the creek, which are macro-properties or surface properties, of the physical systems. Some, but not all, macro-properties can be causally explained by the behaviour of elements at the micro-level. For example, the solidity of the table in front of me is (causally) explained by the lattice structure of the molecules of which the table is composed. Similarly, the liquidity of the water is (causally) explained by the behaviour of the H_2O molecule movements. Not all macro-properties have a causal explanation in terms of micro-behaviour. For example, the velocity of the creek is not explained by the movement of the molecules but rather by the angle of the slope, the pull of gravity and the friction provided by the creek bed. But in the case of those macro-features that are causally explained by the behaviour of elements at the micro-level, it seems to me that we have a perfectly ordinary model for explaining the puzzling relationships between the mind and the brain. In the case of liquidity and solidity, we have no difficulty at all in saying that the surface phenomena are *caused by* the behaviour of elements at the micro-level and at the same time that the surface phenomena *just are* (physical) features of the systems in question. My preferred way of stating this point is to say that the surface feature F is *caused by* the behaviour of micro-elements M, and at the same time is *realized in* the system of micro-elements. The relations between F and M

are causal but at the same time F is simply a higher-level feature of the very system which consists in elements M.

In objection to this one might say F just is, just is identical with, features of M. So, for example, we might define solidity as the lattice structure of the molecular arrangement. This point seems to me correct but not really an objection to the analysis that I am proposing. It is a characteristic of the progress of science that an expression that is originally defined in terms of surface features of a phenomenon, features accessible to the senses, is subsequently defined in terms of the micro-structure that causes the surface features. Thus, to take the example of solidity, the table in front of me is solid in the ordinary sense that it is rigid, it resists pressure, it supports books, it is not easily penetrable by most other objects such as other tables, etc. Such is the common-sense notion of solidity. Now in a scientific vein one can define solidity as whatever micro-structure causes these gross observable features. One can then say either that solidity just is the lattice structure of the system of molecules and that solidity so defined causes, for example, resistance to touch and pressure; or one can say that solidity consists of such things as rigidity and resistance to touch and pressure and is caused by the behaviour of elements at the micro-level. This shift from causation to definitional identity is very common in the history of science. Consider the following pairs: lightning is caused by an electrical discharge – lightning just is an electrical discharge; the colour red is caused by photon emissions with a wavelength of 600 nanometres – red just is a photon emission of 600 nanometres; heat is caused by molecule movements – heat just is the mean kinetic energy of molecule movements.

If we apply these lessons to the study of the mind it seems to me that there is no difficulty in accounting for the metaphysical relations of the mind to the brain in terms of a causal theory of the brain's functioning to produce mental states. Just as the liquidity of water is caused by the behaviour at the micro-level, and yet at the same time is a feature realized in the system of micro-elements, so in exactly that sense of 'caused by' and 'realized in' are mental phenomena caused by processes going on in the brain at the neuronal or modular level but they are realized in the very system that consists of the neurons organized into modules. And, just as we need the micro–macro distinction for any physical system, for the same reasons we need the macro–micro distinction for the brain. Though we can say of a system of particles that it is solid or liquid, we cannot say of any given particle that this particle is solid, or this particle is liquid. In exactly the same way, as far as we know anything at all about it, though we can say of a particular brain that this brain is conscious or this brain is experiencing thirst or pain, we cannot say of any particular neuron that this neuron is in pain, this neuron is experiencing thirst. To repeat this point, though there are enormous empirical mysteries about how the brain works in detail, there are no logical or philosophical or metaphysical obstacles to accounting for the relation between the mind and the brain in terms that are completely familiar to us from the rest of nature. Nothing is more common in nature than for surface features of a phenomenon to be caused by and realized in a micro-structure, and those are exactly the relations that are exhibited by the relation of mind to brain. The

intrinsically *mental* features of the universe are just higher-level *physical* features of brains.

The Possibility of Mental Phenomena

Let us now return to the four problems that seem to face any putative solution to the mind–brain problem.

How is Consciousness Possible?

The best way to show how something is possible is to show how it is actual, and we have already given a sketch of how pains are actually caused by neurophysiological processes going on in the thalamus and the sensory cortex. Why is it then that many people feel dissatisfied with this sort of answer? I think that by pursuing an analogy with an earlier problem in the history of science we can dispel this sense of puzzlement. For a long time many biologists and philosophers thought it impossible, in principle, to account for the phenomena of life on purely biological grounds. They thought that in addition to the biological processes some other element must be necessary, some *élan vital* must be postulated in order to lend life to what was otherwise dead and inert matter. It is hard today to realize how intense the dispute between vitalism and mechanism was even a generation ago, but today these issues are no longer taken seriously. Why not? Is it simply because we synthesized urea (the first organic compound to be synthesized), and this proved that organic compounds could be produced artificially? I think not. I think rather it is because we have come to see the biological character of the processes that are characteristic of living organisms. Once we understand how the features that are characteristic of living beings have a biological explanation, it no longer seems mysterious to us that inert matter should be alive. I think that exactly analogous considerations should apply to our discussions of consciousness. It should seem no more mysterious in principle that this hunk of inert matter, this grey and white oatmeal-textured substance of the brain, should be conscious than it seems to us problematic that this hunk of matter, this collection of nucleic acids, proteins and other molecules stuck on to a calcium frame, should be alive. The way, in short, to dispel the mystery is to understand the processes. We do not yet fully understand the processes, but we understand the *character* of the processes, we understand that there are certain specific electrochemical processes going on in the relations among neurons or neuron-modules and perhaps other features of the brain, and that these processes are causally responsible for the phenomenon of consciousness.

How Can Atoms in the Void Have Intentionality?

As with our first question, the best way to show how something is possible is to show how it is actual. Consider thirst. As far as we know anything about it, at least certain kinds of thirst are caused in the hypothalamus by sequences of neuron firings. These firings are in turn caused by the action of the peptide hormone angiotensin II in the hypothalamus, and

angiotensin II, in turn, is synthesized by renin, which is secreted by the kidneys. Thirst, at least of these kinds, is caused by a series of events in the central nervous system, principally the hypothalamus, and is realized in the hypothalamus. Notice that thirst is an intentional state. To be thirsty is to have, among other things, the desire to drink. Thirst has propositional content, direction of fit, conditions of satisfaction, and all the rest of the features that are common to intentional states.

As with the 'mysteries' of life and consciousness, the way to master the mystery of intentionality is to describe in as much detail as we can how the phenomena are caused by biological processes while at the same time they are realized in biological systems. Visual and auditory experiences, tactile sensations, hunger, thirst, sexual desire and olfactory experiences are all caused by brain processes and realized in the structure of the brain, and all are intentional phenomena. I am not saying we should lose our sense of the mysteries of nature; on the contrary, the examples I have cited are all in a sense astounding. But I am saying that they are neither more nor less mysterious than other astounding features of the world such as the existence of gravitational attraction, the process of photosynthesis or the size of the Milky Way.

Subjectivity

The puzzle about subjectivity can be stated quite simply. Since the seventeenth century our conception of reality has involved the notion of total objectivity. Reality, on this view, is that which is accessible to any competent observer. Indeed, in some versions, reality is that which is objectively measurable. Now the question is: how do we accommodate the subjectivity of mental states within this picture; how do we square the fact that each of us has real subjective mental states with an objectivist conception of the real world? The solution to this puzzle can be stated equally simply. It is a mistake to suppose that the definition of reality should exclude subjectivity. If science is the name of the set of objective and systematic truths we can state about the world, then the existence of subjectivity is just an objective scientific fact like any other. If a scientific account of the world attempts to describe how things are, then one of the features of the account will be the subjectivity of mental states, since it is just a plain fact about biological evolution that it has produced certain sorts of biological systems, namely human and certain animal brains, that have subjective features. My present state of consciousness is a feature of my brain and in consequence is accessible to me in a way that it is not accessible to you, and your present state of consciousness is a feature of your brain and is accessible to you in a way that it is not accessible to me. Thus the existence of subjectivity is an objective physical fact of biology. It is a persistent mistake to try to define 'science' in terms of certain features of existing scientific theories. But once this provincialism is perceived to be the unscientific prejudice it is, then any domain of facts is a subject of scientific investigation. If, for example, God existed, then that fact would be a fact of science like any other. I do not know whether God exists, but I have no doubt at all that subjective mental states exist, because I am now in one and so are you. If the fact of subjectivity runs

counter to a certain definition of 'science', then it is the definition and not the fact which we will have to abandon.

Intentional Causation

The problem of intentional causation for our present purpose is the problem of how to give an account of the mental that avoids epiphenomenalism. How, for example, could anything as gaseous and ethereal as a thought give rise to an action? The answer is that thoughts are not gaseous and ethereal. Their logical and intentional properties are solidly grounded in their causal properties in the brain. Because mental states are physical states of the brain, they can cause behaviour by ordinary causal processes. They have both a higher and a lower level of description, and each level is causally real.

Once again, we can use an analogy from physics to illustrate these relationships. Consider hammering a nail with a hammer. Both hammer and nail must have a certain kind of solidity. Hammers made of cotton or butter will be quite useless, and hammers made of water or steam are not hammers at all. Solidity is a real causal property of the hammer and not something epiphenomenal. But the solidity itself is caused by the behaviour of particles at the micro-level and is realized in the system of micro-elements. The existence of two causally real levels of description in the brain, one a macro-level of mental neurophysiological processes and the other a micro-level of neuronal physiological processes, is exactly analogous to the existence of two causally real levels of description of the hammer. Consciousness, for example, is a real causal property of the brain and not something epiphenomenal. My conscious attempt to perform an action such as raising my arm causes the movement of the arm. At the higher level of description, the intention to raise my arm has the movement of my arm as its condition of satisfaction and it causes the movement of the arm. At the lower level of description, a series of neuron firings which originate in the cortex causes the release of the transmitter substance acetylcholine at the 'end plates' where the axon terminals of motor neurons connect to the muscle fibres; this in turn causes a series of chemical changes that result in the contraction of the muscle. As with the case of hammering a nail, the same sequence of events has two levels of description, both of which are causally real and where the higher-level causal features are both caused by and realized in the structure of the lower-level elements.

Traditional Categories

I have so far resisted using the traditional vocabulary of dualism, monism, physicalism, etc. in attempting to characterize the view argued for in this chapter. However, it may be useful to see how these views relate to the traditional categories. In a discussion of these matters at the Philosophy of Mind conference at New York University, Hilary Putnam from Harvard characterized the view put forward here as (1) property dualism; (2) emergentism; (3) supervenience. I think it will deepen our understanding of the issues if we consider each of these assessments in turn.

Property Dualism

If 'property dualism' is simply the view that the world contains some physical features which are mental – my present state of consciousness, for example – and some physical features which are non-mental – the weight of my brain, for example – then my view can correctly be described as property dualism. Nonetheless, I believe that there is something deeply misleading about this characterization. 'Property dualism' seems to imply that there are two and only two types of properties in the world, physical and mental, and that is emphatically not the view that I hold. On my view, mental properties just are higher-level physical features of certain physical systems in the same sense that solidity and liquidity are higher-level physical features of certain physical systems. Thus mental properties are physical properties in the sense that liquidity and solidity are physical properties. This view, it seems to me, is correctly described not so much as property dualism, but as *property polyism*. That is, there are lots of different kinds of higher-level properties of systems, and mental properties are among them. To put this point another way, on my view the words 'mental' and 'physical' are not properly opposed to each other, because mental properties, naively construed, are just one class of physical properties, and physical properties are correctly opposed not to mental properties but to such other features as logical properties and ethical properties, for example.

Emergentism

Similar considerations apply to the question whether or not we should think of mental properties as in some way emergent. It all depends on what you mean by 'emergent'. If we are to think of any higher-level feature of a system such as solidity, liquidity, etc. as emergent, then in that sense I believe states of consciousness, intentionality, subjectivity, etc. are indeed emergent properties of certain biological systems. In fact, if we define emergent properties of a system of elements as properties which can be explained by the behaviour of the individual elements but which are not properties of elements construed individually, then it is a trivial consequence of my view that mental properties are emergent properties of neurophysiological systems. However, traditionally, emergentism is often regarded as implying something mysterious; it is taken as implying that there is some mysterious non-physical process that produces a peculiar kind of property. Emergentism, in short, tends to go with the more mysterious aspects of dualism, and in that sense I am denying that my view can correctly be characterized as emergentism. If emergentism is taken to imply that there is something mysterious, something lying outside the scope of physical or biological sciences as they are normally construed, in the existence of emergent properties, then it seems to me clear that mental properties are not emergent in that sense.

Supervenience

The doctrine of the supervenience of the mental on the physical is the doctrine that there can be no mental differences without corresponding

physical differences: if a system is in two different mental states at two different times, then it must have different physical properties at those two times. This view is a consequence of the thesis that mental phenomena are caused by and realized in the brain, for if the effects are different the causes must be different. Indeed, it seems to me a merit of the view advanced here that the supervenience of the mental is simply a special case of the general principle of the supervenience of macro-properties on micro-properties. There is nothing special or arbitrary or mysterious about the supervenience of the mental on the physical; it is simply one more instance of the supervenience of higher-order physical properties on lower-order physical properties. If a bowl of water is ice at one time and liquid later, then there must be a difference in the behaviour of the micro-particles to account for this difference. Similarly, if I want a drink of water at one time and later do not want a drink of water, there must be a difference in my brain to account for this difference in my mental states.

Consequences for the Philosophy of Mind

Some mental concepts, such as, for example, *having a pain* or *believing* that so and so, denote entities that exist entirely in the mind. Others such as *seeing* or *knowing* also refer to mental phenomena but they require that additional conditions be met in order that the concept be applicable. So, for example, to say that X knows that P implies more than that X believes that P; it implies, among other things that P is true, and the truth of P cannot, in general, be solely a matter of what goes on in the mind of X. To say that X sees that P implies that X has a visual experience of a certain sort, but it also implies that it is the case that P. Let us call such concepts whose truth conditions depend only on what goes on in the mind 'pure mental concepts' and let us call such mental concepts as those whose truth conditions require extra-mental phenomena 'hybrid mental concepts'. Now since hybrid mental concepts all contain by definition a mental component, to the extent that we are discussing the nature of the mind, we can carve off that mental component and examine it separately. For every hybrid mental concept there is a corresponding pure mental concept which captures the purely mental component of the hybrid concept. As far as the mind proper is concerned, we can confine our discussion entirely to pure mental concepts and the pure mental states which are the denotations of pure mental concepts. Whenever a mental phenomenon is present in the mind of an agent – for example, he is feeling a pain, thinking about philosophy or wishing he had a cold beer – causally sufficient conditions for that phenomenon are entirely in the brain. And indeed the thesis that mental phenomena are caused by and realized in the brain has the consequence that, for any mental phenomenon whatever, causally sufficient conditions are in the brain. Let us call this principle *the principle of neurophysiological sufficiency*. Now if this principle is true, then many current theories in the philosophy of mind will turn out to be false, because they are inconsistent with this principle. For example, several philosophers following Wittgenstein and Heidegger have tried to explain the intentionality of mental phenomena in terms of

social relationships. But how are we to take this explanation? If we take it as claiming that social relationships are necessary for, or constitutive of, mental life, then we know that it must be false, because social relationships are relevant to the causal production of intentionality only in so far as they impact on the brains of human agents; and the actual mental states, beliefs, desires, hopes, fears and the rest of it have causally sufficient conditions that are entirely internal to the nervous system. This is not to deny that social relations are crucial for the production of many forms of intentionality such as, for example, language. Children can learn

Figure 15.5 Ludwig Wittgenstein (1889–1951).

a language and use it only if they are exposed to other people who also use language. But the thesis that there are forms of intentionality that require a social base needs to be interpreted so that it is consistent with the claim that intentionality is a purely internal product of internal physiological processes. These views are not necessarily inconsistent; they can be interpreted as just ways of describing different aspects of the same phenomenon. The mistake is to suppose that the social relations can somehow or other replace or substitute for what goes on in the brain.

An even more prominent implicit denial of the principle of neurophysiological sufficiency is the entire tradition that has been built around Wittgenstein's claim that 'an inner process stands in need of an outward criterion'.[15] So, for example, Norman Malcolm has tried to give a non-internal account of dreaming,[16] Elizabeth Anscombe has tried to explain intentions in terms of outward behaviour,[17] and Anthony Kenny has tried to explain many emotions in terms of their social setting and behavioural consequences.[18] But it is hard to interpret any of these analyses in ways that are consistent with the principle of neurophysiological sufficiency. Whatever other features dreams have, they are caused by neurophysiological processes. And the same goes for intentions and emotions such as fear and anger. Now perhaps we might interpret the views of Kenny, Malcolm and Anscombe as simply describing constraints on our having a *vocabulary* for discussing mental phenomena. And perhaps we might interpret Wittgenstein's claim as the claim that a *vocabulary* for inner processes stands in need of outward criteria. But if we take these claims as claims about the *nature* of the mental phenomena themselves – that one can't have a dream or have an intention or be angry except when certain external conditions are satisfied, conditions external to the brain, that is – then we know these theses must be false because of the principle of neurophysiological sufficiency. What goes on in the head must be causally sufficient for any mental state whatever.

And, of course, this Wittgensteinian tradition is itself part of a larger tradition of seeking behaviouralistic or quasi-behaviouralistic analyses of mental concepts. And once again, we know from the principle of neurophysiological sufficiency that these efforts are doomed to failure. We cannot define mental phenomena in terms of their behavioural manifestations, because we know that it is always possible to have the phenomena independently of having any behavioural manifestations.

Some Conclusions

The main polemical aims of this chapter regarding the relations of minds and programs can be swiftly summarized. In order that there be total clarity, I will state a set of 'axioms' and derive the relevant conclusions.

Axiom 1. Brains cause minds

This is simply a very crude statement of the empirical fact that the relevant causal processes in the brain are sufficient to produce any mental phenomenon. It is important to re-emphasize that, where pure mental phenomena are concerned, there is no essential connection between these internal causal processes that are sufficient for mental phenomena and the causal input–output relations of the whole system. In principle, we could have all of our mental life without any of the appropriate stimuli or any of the normal external behaviour.

Axiom 2. Syntax is not sufficient for semantics

This is a conceptual or logical truth that articulates the distinction between the level of formal symbols and the level of meaning.

Axiom 3. Minds have contents; specifically, they have intentional or semantic contents

Axiom 4. Programs are defined purely formally, or syntactically

Now from these obvious points we can derive some controversial conclusions.

Conclusion 1. Instantiating a program by itself is never sufficient for having a mind (by Axioms 2, 3 and 4)

This conclusion by itself is sufficient to refute Strong Artificial Intelligence.

Conclusion 2. The way the brain functions to cause minds cannot be solely by instantiating a program (Axiom 1 and Conclusion 1)

Conclusion 3. Any artefact that had a mind would have to have causal powers (at least) equivalent to those of the brain (by Axiom 1, trivially)

Conclusion 4. For any artefact that had a mind, the program by itself would not be sufficient for having a mind. The artefact would have to have causal powers equivalent to the brain (by Conclusions 1 and 3)

Anyone who wishes to challenge the central theses owes us a precise specification of which 'axioms' and which derivations are being challenged.

Notes and References

1 This chapter is based on a lecture delivered in Oxford at a time when I was in Britain for the purpose of taping the 1984 Reith Lectures for the BBC, and it was not originally intended for separate publication. There is considerable overlap of the material in the Reith Lectures and the material in this lecture. The lectures have since been published as *Minds, Brains and Science* (BBC Publications, 1984; Harvard Univ. Press, 1984). I apologize to the listeners and readers of the Reith Lectures for the repetition. I am publishing this article separately, in part because Colin Blakemore and Susan Greenfield have convinced me that it might make a useful contribution to the volume, in spite of the repetition of material published elsewhere, and in part because it gives me the opportunity to expand and explain further several of the points that were made in the Reith Lectures.

2 McCarthy, John (1979) Ascribing mental qualities to machines. Stanford Artificial Intelligence Laboratory Memo AIM-326, p. 2. *Computer Science Department Report*, no. STAN-CS-79-725, March 1979.

3 Schank, R. C. and Abelson, R. P. (1977) *Scripts, Plans, Goals and Understanding.* Hillsdale, NJ: Erlbaum.

4 I do not mean to imply that either Schank or Abelson makes this claim.

5 'Language understanding' programs characteristically have a 'syntax', a 'semantics', and in some cases even a 'pragmatics'. This is, of course, quite irrelevant to the present point, since all three levels are computational, i.e. 'syntactical' in the sense in which I am now using the word.

6 Some truly desperate people in AI have proposed that it is not *I* who

understand but the *whole room*; i.e. the system containing me, the program, the bushel baskets, the window to the outside, etc. But this reply is subject to exactly the same objection. Just as I do not have any way of getting from syntax to semantics, neither does the whole system. The whole system has no way of knowing what any of these formal symbols actually *mean*.

7 Ringle, Martin (1980) Mysticism as a philosophy of artificial intelligence. Commentary on Searle's 'Minds, brains, and programs'. *The Behavioral and Brain Sciences*, 3: 444.

8 Dennett, Daniel (1980) The milk of human intentionality. Commentary on Searle's 'Minds, brains, and programs'. *The Behavioral and Brain Sciences*, 3: 428.

9 Hofstadter, Douglas R. (1980) Reductionism and religion. Commentary on Searle's 'Minds, brains, and programs'. *The Behavioral and Brain Sciences*, 3: 433.

10 Hubel, D. (1978) Vision and the brain. *Bulletin of the American Academy of Arts and Sciences*, April 1978, 31, no. 7: 18.

11 Oppenheim, Paul and Putnam, Hilary (1958) Unity of science as a working hypothesis. In Feigl, Scriven and Maxwell (eds), *Minnesota Studies in the Philosophy of Science*, vol. 2, Concepts, Theories, and the Mind–Body Problem, p. 19. Minneapolis: Univ. of Minnesota Press.

12 See Szentagothai's chapter in the book.

13 Edelman, Gerald (1982) Through a computer darkly: group selection and higher brain function. *Bulletin of the American Academy of Arts and Sciences*, October 1982, 36, no. 1: 28.

14 Phantom limb pains are pains suffered by amputees which feel as if they were coming from the, now non-existent, limb.

15 Wittgenstein, Ludwig (1973) *Philosophical Investigations*, tr. G. E. M. Anscombe. New York: Macmillan.

16 Malcolm, Norman (1959) *Dreaming*. London: Routledge & Kegan Paul.

17 Anscombe, G. E. M. (1963) *Intention*. Ithaca, NY: Cornell Univ. Press.

18 Kenny, Anthony (1963) *Action, Emotion and Will*. London: Routledge & Kegan Paul.