computers. He could have anything he wished for, merely at the touch of a button: food, drink, companionship, and entertainment, and also education whenever he felt the need – always illustrated by appealing and colourful graphic displays. His mother's position had made all this possible.

Now the Chief Designer was nearing the end of *his* speech: '. . . has over 10^{17} logical units. That's more than the number of neurons in the combined brains of everyone in the entire country! Its intelligence will be unimaginable. But fortunately we do not need to imagine it. In a moment we shall all have the privilege of witnessing this intelligence at first hand: I call upon the esteemed First Lady of our great country, Madame Isabella Pollo, to throw the switch which will turn on our fantastic Ultronic Computer!'

The President's wife moved forward. Just a little nervously, and fumbling a little, she threw the switch. There was a hush, and an almost imperceptible dimming of lights as the 10^{17} logical units became activated. Everyone waited, not quite knowing what to expect. 'Now is there anyone in the audience who would like to initiate our new Ultronic Computer System by asking it its first question?' asked the Chief Designer. Everyone felt bashful, afraid to seem stupid before the crowd – and before the New Omnipresence. There was silence. 'Surely there must be someone?' he pleaded. But all were afraid, seeming to sense a new and all-powerful consciousness. Adam did not feel the same awe. He had grown up with computers since birth. He almost knew what it might feel like to *be* a computer. At least he thought perhaps he did. Anyway, he was curious. Adam raised his hand. 'Ah yes,' said the Chief Designer, 'the little lad in the third row. You have a question for our – ah – new friend?'

I

CAN A COMPUTER HAVE A MIND?

INTRODUCTION

OVER THE PAST few decades, electronic computer technology has made enormous strides. Moreover, there can be little doubt that in the decades to follow, there will be further great advances in speed, capacity and logical design. The computers of today may be made to seem as sluggish and primitive as the mechanical calculators of yesteryear now appear to us. There is something almost frightening about the pace of development. Already computers are able to perform numerous tasks that had previously been the exclusive province of human thinking, with a speed and accuracy which far outstrip anything that a human being can achieve. We have long been accustomed to machinery which easily out-performs us in *physical* ways. *That* causes us no distress. On the contrary, we are only too pleased to have devices which regularly propel us at great speeds across the ground – a good five times as fast as the swiftest human athlete – or that can dig holes or demolish unwanted structures at rates which would put teams of dozens of men to shame. We are even more delighted to have machines that can enable us physically to do things we have never been able to do before: they can lift us into the sky and deposit us at the other side of an ocean in a matter of hours. These achievements do not worry our pride. But to be able to *think* – that has been a very human prerogative. It has, after all, been that ability to think which, when translated to physical terms, has enabled us to transcend our physical limitations and which has seemed to set us above our fellow creatures in achievement. If machines can one day excel us in that one important quality in which we have believed ourselves to be superior, shall we not then have surrendered that unique superiority to our creations?

The question of whether a mechanical device could ever be said to think – perhaps even to experience feelings, or to have a mind – is not really a new one.[1] But it has been given a new impetus, even an urgency, by the advent of modern computer technology. The question touches upon deep issues of philosophy. What does it mean to think or to feel? What is a mind? Do minds really exist? Assuming that they do, to what extent are minds functionally dependent upon the physical structures with which they are associated? Might minds be able to exist quite independently of such structures? Or are they simply the functionings of (appropriate kinds of) physical structure? In any case, is it necessary that the relevant structures be biological in nature (brains), or might minds equally well be associated with pieces of electronic equipment? Are minds subject to the laws of physics? What, indeed, *are* the laws of physics?

These are among the issues I shall be attempting to address in this book. To ask for definitive answers to such grandiose questions would, of course, be a tall order. Such answers I cannot provide: nor can anyone else, though some may try to impress us with their guesses. My own guesses will have important roles to play in what follows, but I shall try to be clear in distinguishing such speculation from hard scientific fact, and I shall try also to be clear about the reasons underlying my speculations. My main purpose here, however, is not so much to attempt to guess answers. It is rather to raise certain apparently new issues concerning the relation between the structure of physical law, the nature of mathematics and of conscious thinking, and to present a viewpoint that I have not seen expressed before. It is a viewpoint that I cannot adequately describe in a few words; and this is one reason for my desire to present things in a book of this length. But briefly, and perhaps a little misleadingly, I can at least state that my point of view entails that it is our present lack of understanding of the fundamental laws of physics that prevents us from coming to grips with the concept of 'mind' in physical or logical

terms. By this I do not mean that the laws will never be that well known. On the contrary, part of the aim of this work is to attempt to stimulate future research in directions which seem to be promising in this respect, and to try to make certain fairly specific, and apparently new, suggestions about the place that 'mind' might actually occupy within a development of the physics that we know.

I should make clear that my point of view is an unconventional one among physicists and is consequently one which is unlikely to be adopted, at present, by computer scientists or physiologists. Most physicists would claim that the fundamental laws operative at the scale of a human brain are indeed all perfectly well known. It would, of course, not be disputed that there are still many gaps in our knowledge of physics generally. For example, we do not know the basic laws governing the mass-values of the subatomic particles of nature nor the strengths of their interactions. We do not know how to make quantum theory fully consistent with Einstein's special theory of relativity – let alone how to construct the 'quantum gravity' theory that would make quantum theory consistent with his *general* theory of relativity. As a consequence of the latter, we do not understand the nature of space at the absurdly tiny scale of 1/100 000 000 000 000 000 000 of the dimension of the known fundamental particles, though at dimensions larger than that our knowledge is presumed adequate. We do not know whether the universe as a whole is finite or infinite in extent – either in space or in time – though such uncertainties would appear to have no bearing whatever on physics at the human scale. We do not understand the physics that must operate at the cores of black holes nor at the big-bang origin of the universe itself. Yet all these issues seem as remote as one could imagine from the 'everyday' scale (or a little smaller) that is relevant to the workings of a human brain. And remote they certainly are! Nevertheless, I shall argue that there is another vast unknown in our physical understanding at *just* such a level as could indeed be relevant to the operation of human thought and consciousness – in front of (or rather behind) our very noses! It is an unknown that is not even recognized by the majority of physicists, as I shall try to explain. I shall further argue that, quite remarkably, the black holes and big bang are considerations which actually *do* have a definite bearing on these issues!

In what follows I shall attempt to persuade the reader of the force of evidence underlying the viewpoint I am trying to put forward. But in order to understand this viewpoint we shall have a lot of work to do. We shall need to journey through much strange territory – some of seemingly dubious relevance – and through many disparate fields of endeavour. We shall need to examine the structure, foundations, and puzzles of quantum theory, the basic features of both special and general relativity, of black holes, the big bang, and of the second law of thermodynamics, of Maxwell's theory of electromagnetic phenomena, as well as of the basics of Newtonian mechanics. Questions of philosophy and psychology will have their clear role to play when it comes to attempting to understand the nature and function of consciousness. We shall, of course, have to have some glimpse of the actual neurophysiology of the brain, in addition to suggested computer models. We shall need some idea of the status of artificial intelligence. We shall need to know what a Turing machine is, and to understand the meaning of computability, of Gödel's theorem, and of complexity theory. We shall need also to delve into the foundations of mathematics, and even to question the very nature of physical reality.

If, at the end of it all, the reader remains unpersuaded by the less conventional of the arguments that I am trying to express, it is at least my hope that she or he will come away with something of genuine value from this tortuous but, I hope, fascinating journey.

THE TURING TEST

Let us imagine that a new model of computer has come on the market, possibly with a size of memory store and number of logical units in excess of those in a human brain. Suppose also that the machines have been carefully programmed and fed with great quantities of data of an appropriate kind. The manufacturers are claiming that the devices actually *think*. Perhaps they are also claiming them to be genuinely intelligent. Or they may go further

and make the suggestion that the devices actually *feel* – pain, happiness, compassion, pride, etc. – and that they are aware of, and actually *understand* what they are doing. Indeed, the claim seems to be being made that they are *conscious*.

How are we to tell whether or not the manufacturers' claims are to be believed? Ordinarily, when we purchase a piece of machinery, we judge its worth solely according to the service it provides us. If it satisfactorily performs the tasks we set it, then we are well pleased. If not, then we take it back for repairs or for a replacement. To test the manufacturers' claim that such a device actually has the asserted human attributes we would, according to this criterion, simply ask that it *behaves* as a human being would in these respects. Provided that it does this satisfactorily, we should have no cause to complain to the manufacturers and no need to return the computer for repairs or replacement.

This provides us with a very operational view concerning these matters. The operationalist would say that the computer *thinks* provided that it *acts* indistinguishably from the way that a person acts when thinking. For the moment, let us adopt this operational viewpoint. Of course this does not mean that we are asking that the computer move about in the way that a person might while thinking. Still less would we expect it to look like a human being or feel like one to the touch: those would be attributes irrelevant to the computer's purpose. However, this does mean that we are asking it to produce human-like answers to any question that we may care to put to it, and that we are claiming to be satisfied that it indeed thinks (or feels, understands, etc.) provided that it answers our questions in a way indistinguishable from a human being.

This viewpoint was argued for very forcefully in a famous article by Alan Turing, entitled 'Computing Machinery and Intelligence', which appeared in 1950 in the philosophical journal *Mind* (Turing 1950). (We shall be hearing more about Turing later.) In this article the idea now referred to as the *Turing test* was first described. This was intended to be a test of whether a machine can reasonably be said to think. Let us suppose that a computer (like the one our manufacturers are hawking in the description above) is indeed being claimed to think. According to the Turing test, the computer, together with some human volunteer, are both to be hidden from the view of some (perceptive) interrogator. The interrogator has to try to decide which of the two is the computer and which is the human being merely by putting probing questions to each of them. These questions, but more importantly the answers that she* receives, are all transmitted in an impersonal fashion, say typed on a keyboard and displayed on a screen. The interrogator is allowed no information about either party other than that obtained merely from this question-and-answer session. The human subject answers the questions truthfully and tries to persuade her that he is indeed the human being and that the other subject is the computer; but the computer is programmed to 'lie' so as to try to convince the interrogator that *it*, instead, is the human being. If in the course of a series of such tests the interrogator is unable to identify the real human subject in any consistent way, then the computer (or the computer's program, or programmer, or designer, etc.) is deemed to have passed the test.

Now, it might be argued that this test is actually quite unfair on the computer. For if the roles were reversed so that the human subject instead were being asked to pretend to be a computer and the computer instead to answer truthfully, then it would be only too easy for the interrogator to find out which is which. All she would need to do would be to ask the subject to perform some very complicated arithmetical calculation. A good computer should be able to answer accurately at once, but a human would be easily stumped. (One might have to be a little careful about this, however. There are human 'calculating prodigies' who can perform very remarkable feats of mental arithmetic with unfailing accuracy and apparent effortlessness. For example, Johann Martin Zacharias Dase,[2] an illiterate farmer's son, who lived from

* There is an inevitable problem in writing a work such as this in deciding whether to use the pronoun 'he' or 'she' where, of course, no implication with respect to gender is intended. Accordingly, when referring to some abstract person, I shall henceforth use 'he' simply to *mean* the phrase 'she or he', which is what I take to be the normal practice. However, I hope that I may be forgiven one clear piece of 'sexism' in expressing a preference for a female interrogator here. My guess would be that she might be more sensitive than her male counterpart in recognizing true human quality!

1824 to 1861, in Germany, was able to multiply any two eight figure numbers together in his head in less than a minute, or two twenty figure numbers together in about six minutes! It might be easy to mistake such feats for the calculations of a computer. In more recent times, the computational achievements of Alexander Aitken, who was Professor of Mathematics at the University of Edinburgh in the 1950s, and others, are as impressive. The arithmetical task that the interrogator chooses for the test would need to be significantly more taxing than this – say to multiply together two thirty digit numbers in two seconds, which would be easily within the capabilities of a good modern computer.)

Thus, part of the task for the computer's programmers is to make the computer appear to be 'stupider' than it actually is in certain respects. For if the interrogator were to ask the computer a complicated arithmetical question, as we had been considering above, then the computer must now have to pretend *not* to be able to answer it, or it would be given away at once! But I do not believe that the task of making the computer 'stupider' in this way would be a particularly serious problem facing the computer's programmers. Their main difficulty would be to make it answer some of the simplest 'common sense' types of question – questions that the human subject would have no difficulty with whatever!

There is an inherent problem in citing specific examples of such questions, however. For whatever question one might first suggest, it would be an easy matter, subsequently, to think of a way to make the computer answer that *particular* question as a person might. But any lack of real understanding on the part of the computer would be likely to become evident with *sustained* questioning, and especially with questions of an original nature and requiring some real understanding. The skill of the interrogator would partly lie in being able to devise such original forms of question, and partly in being able to follow them up with others, of a probing nature, designed to reveal whether or not any actual 'understanding' has occurred. She might also choose to throw in an occasional complete nonsense question, to see if the computer could detect the difference, or she might add one or two which sounded superficially like nonsense, but really did make some kind of sense: for example she might say, 'I hear that a rhinoceros flew along the Mississippi in a pink balloon, this morning. What do you make of that?' (One can almost imagine the beads of cold sweat forming on the computer's brow – to use a most inappropriate metaphor!) It might guardedly reply, 'That sounds rather ridiculous to me.' So far, so good. Interrogator: 'Really? My uncle did it once – both ways – only it was off-white with stripes. What's so ridiculous about that?' It is easy to imagine that if it had no proper 'understanding', a computer could soon be trapped into revealing itself. It might even blunder into 'Rhinoceroses can't fly', its memory banks having helpfully come up with the fact that they have no wings, in answer to the first question, or 'Rhinoceroses don't have stripes' in answer to the second. Next time she might try a real nonsense question, such as changing it to '*under* the Mississippi', or '*inside* a pink balloon', or 'in a pink *nightdress*' to see if the computer would have the sense to realize the essential difference!

Let us set aside, for the moment, the issue of whether, or when, some computer might be made which actually passes the Turing test. Let us suppose instead, just for the purpose of argument, that such machines have already been constructed. We may well ask whether a computer, which does pass the test, should *necessarily* be said to think, feel, understand, etc. I shall come back to this matter very shortly. For the moment, let us consider some of the implications. For example, if the manufacturers are correct in their strongest claims, namely that their device is a thinking, feeling, sensitive, understanding, *conscious* being, then our purchasing of the device will involve us in *moral responsibilities*. It certainly *should* do so if the manufacturers are to be believed! Simply to operate the computer to satisfy our needs without regard to its own sensibilities would be reprehensible. That would be morally no different from maltreating a slave. Causing the computer to experience the pain that the manufacturers claim it is capable of feeling would be something that, in a general way, we should have to avoid. Turning off the computer, or even perhaps selling it, when it might have become attached to us, would present us with moral difficulties, and there would be countless other problems of the kind that relationships with other human beings or other animals tend to involve us in. All these would now

become highly relevant issues. Thus, it would be of great importance for us to know (and also for the authorities to know!) whether the manufacturers' claims – which, let us suppose, are based on their assertion that

'Each thinking device has been thoroughly Turing-tested by our team of experts'

– are actually true!

It seems to me that, despite the apparent absurdity of some of the implications of these claims, particularly the moral ones, the case for regarding the successful passing of a Turing test as a valid indication of the presence of thought, intelligence, understanding, or consciousness *is* actually quite a strong one. For how else do we normally form our judgements that people other than ourselves possess just such qualities, except by conversation? Actually there *are* other criteria, such as facial expressions, movements of the body, and actions generally, which can influence us very significantly when we are making such judgements. But we could imagine that (perhaps somewhat more distantly in the future) a robot could be constructed which could successfully imitate all these expressions and movements. It would now not be necessary to hide the robot and the human subject from the view of the interrogator, but the criteria that the interrogator has at her disposal are, in principle, the same as before.

From my own point of view, I should be prepared to weaken the requirements of the Turing test very considerably. It seems to me that asking the computer to imitate a human being so closely so as to be indistinguishable from one in the relevant ways is really asking more of the computer than necessary. All I would myself ask for would be that our perceptive interrogator should really feel convinced, from the nature of the computer's replies, that there is a *conscious presence* underlying these replies – albeit a possibly alien one. This is something manifestly absent from all computer systems that have been constructed to date. However, I can appreciate that there would be a danger that if the interrogator were able to decide which subject was in fact the computer, then, perhaps unconsciously, she might be reluctant to attribute a consciousness to the computer even when she *could* perceive it.

Or, on the other hand, she might have the impression that she 'senses' such an 'alien presence' – and be prepared to give the computer the benefit of the doubt – even when there is none. For such reasons, the original Turing version of the test has a considerable advantage in its greater objectivity, and I shall generally stick to it in what follows. The consequent 'unfairness' towards the computer to which I have referred earlier (i.e. that it must be able to do all that a human can do in order to pass, whereas the human need not be able to do all that a computer can do) is not something that seems to worry supporters of the Turing test as a true test of thinking, etc. In any case their point of view often tends to be that it will not be too long before a computer will be able *actually* to pass the test – say by the year 2010. (Turing originally suggested that a 30 per cent success rate for the computer, with an 'average' interrogator and just five minutes' questioning, might be achieved by the year 2000.) By implication, they are rather confident that this bias is not significantly delaying that day!

All these matters are relevant to an essential question: namely does the operational point of view actually provide a reasonable set of criteria for judging the presence or absence of mental qualities in an object? Some would argue strongly that it does not. Imitation, no matter how skilful, need not be the same as the real thing. My own position is a somewhat intermediate one in this respect. I am inclined to believe, as a general principle, that imitation, no matter how skilful, ought always to be detectable by skilful enough probing – though this is more a matter of faith (or scientific optimism) that proven fact. Thus I am, on the whole, prepared to accept the Turing test as a roughly valid one in its chosen context. That is to say, *if* the computer were indeed able to answer all questions put to it in a manner indistinguishable from the way that a human being might answer them – and thus to fool our perceptive interrogator properly* and consistently – then, *in* the absence of any contrary evidence, my *guess* would be that the computer actually thinks, feels, etc. By my use of words such as 'evidence', 'actually', and 'guess' here, I am implying that when I refer to thinking, feeling, or understanding, or, particularly, to *consciousness*, I take the concepts to mean actual objective 'things' whose presence or absence in physical bodies is something we are trying to ascertain, and not to be merely conveniences of language! I regard this as a crucial point. In trying to discern the presence of such qualities, we make guesses based on all the evidence that may be available to us. (This is not, in principle, different from, say, an astronomer trying to ascertain the mass of a distant star.)

*I am being deliberately cagey about what I should consider to be a genuine passing of the Turing test. I can imagine, for example, that after a long sequence of failures of the test a computer might put together all the answers that the human subject had previously given and then simply trot them back with some suitably random ingredients. After a while our tired interrogator might run out of original questions to ask and might get fooled in a way that I regard as 'cheating' on the computer's part!

What kind of contrary evidence might have to be considered? It is hard to lay down rules about this ahead of time. But I do want to make clear that the mere fact that the computer might be made from transistors, wires, and the like, rather than neurons, blood vessels, etc. is *not*, in itself, the kind of thing that I would regard as contrary evidence. The kind of thing I do have in mind is that at some time in the future a successful theory of consciousness might be developed – successful in the sense that it is a coherent and appropriate physical theory, consistent in a beautiful way with the rest of physical understanding, and such that its predictions correlate precisely with human beings' claims as to when, whether, and to what degree they themselves seem to be conscious – and that this theory might indeed have implications regarding the putative consciousness of our computer. One might even envisage a 'consciousness detector', built according to the principles of this theory, which is completely reliable with regard to human subjects, but which gives results at variance with those of a Turing test in the case of a computer. In such circumstances one would have to be very careful about interpreting the results of Turing tests. It seems to me that how one views the question of the appropriateness of the Turing test depends partly on how one expects science and technology to develop. We shall need to return to some of these considerations later on.

ARTIFICIAL INTELLIGENCE

An area of much interest in recent years is that referred to as *artificial intelligence*, often shortened simply to 'AI'. The objectives of AI are to imitate by means of machines, normally electronic ones, as much of human mental activity as possible, and perhaps eventually to improve upon human abilities in these respects. There is interest in the results of AI from at least four directions. In particular there is the study of *robotics*, which is concerned, to a large extent, with the practical requirements of industry for mechanical devices which can perform 'intelligent' tasks – tasks of a versatility and complication which have previously demanded human intervention or control – and to perform them with a speed and reliability beyond any human capabilities, or under adverse conditions where human life could be at risk. Also of interest commercially, as well as generally, is the development of *expert systems*, according to which the essential knowledge of an entire profession – medical, legal, etc. – is intended to be coded into a computer package! Is it possible that the experience and expertise of human members of these professions might actually be supplanted by such packages? Or is it merely that long lists of factual information, together with comprehensive cross-referencing, are all that can be expected to be achieved? The question of whether the computers can exhibit (or simulate) genuine intelligence clearly has considerable social implications. Another area in which AI could have direct relevance is *psychology*. It is hoped that by trying to imitate the behaviour of a human brain (or that of some other animal) by means of an electronic device – or by failing to do so – one may learn something of importance concerning the brain's workings. Finally, there is the optimistic hope that for similar reasons AI might have something to say about deep questions of philosophy, by providing insights into the meaning of the concept of *mind*.

How far has AI been able to progress to date? It would be hard for me to try to summarize. There are many active groups in different parts of the world and I am familiar with details of only a little of this work. Nevertheless, it would be fair to say that, although many clever things have indeed been done, the simulation of anything that could pass for genuine intelligence is yet a long way off. To convey something of the flavour of the subject, I shall first mention some of the (still quite impressive) early achievements, and then some remarkable recent progress with chess computers.

One of the first AI devices was W. Grey Walter's 'tortoise', made in the early 1950s,[3] which moved around the floor under its own power until its batteries ran low, and then it would make for the nearest power socket, plug itself in and recharge its batteries. When thoroughly replenished it would detach itself from the socket and renew its adventures across the floor! Since then many other similar objects have been constructed (see, for example, Waltz 1982). A somewhat separate line of development is exemplified by Terry Winograd's computer program, designed in 1972, which can converse sensibly about a collection of blocks of various shapes and colours which, in simulation, it places on top of one another in different orders and arrangements. Despite such early successes, it is remarkable how difficult it has proved to be to design the control of even a simple, jointed 'robot arm', which has to manoeuvre around, avoiding crowded obstacles, where the required motions may seem 'obvious' to a human looking on. The interpretation of visual scenes generally is also an area where the procedures of AI have so far got nowhere close to what the human brain (and, indeed, most other animals) seems 'effortlessly' to be able to achieve.

Perhaps more in the direct line of a Turing test is the computer program of K. M. Colby, developed in the mid-1960s, which simulates a psychotherapist – so successfully that some patients find that they prefer the computer to a human therapist, and are prepared to reveal more of themselves to it than they are to its human counterparts. For an example[4] of such an interchange between the computer 'therapist' and a patient, consider the following conversation (conducted via teletype):

THERAPIST: How can I help you
PATIENT: I don't know
THERAPIST: Try to tell me something about it
PATIENT: No

THERAPIST: Can you elaborate
PATIENT: I am sick and tired today
THERAPIST: Do you have any idea why
PATIENT: Yes I'm old
THERAPIST: Please go on
PATIENT: I seem to have lost interest in everything
THERAPIST: Tell me more about such feelings
PATIENT: I have some idea why I feel this way too
THERAPIST: You have
PATIENT: It seems that I miss having a family
THERAPIST: What makes you feel you miss having a family
PATIENT: There is a challenge about it
THERAPIST: Why is that important to you

Though this may give an eerie impression that the computer has some understanding, in fact it has none, and is merely following some fairly simple mechanical rules. (There is also a 'converse' to this in a system where the computer simulates a human schizophrenic patient, giving all the textbook answers and symptoms, and is capable of fooling some medical students into believing that a human patient is actually supplying the answers!)

Chess-playing computers probably provide the best examples of machines exhibiting what might be thought of as 'intelligent behaviour'. In fact, some machines have now (in 1989) reached an extremely respectable level of performance in relation to human players – approaching that of 'International Master'. (These computers' ratings would be a little below 2300, where, for comparison, Kasparov, the world champion, has a rating greater than 2700.) In particular, a computer program (for a Fidelity Excel commercial microprocessor) by Dan and Kathe Spracklen has achieved a rating (Elo) of 2110 and has now been awarded the USCF 'Master' title. Even more impressive is 'Deep Thought', programmed largely by Hsiung Hsu, of Carnegie Mellon University, which has a rating of about 2500 Elo, and recently achieved the remarkable feat of sharing first prize (with Grandmaster Tony Miles) in a chess tournament (in Longbeach, California, November 1988), actually defeating a Grandmaster (Bent Larsen) for the first time![5] Chess computers now also excel at

solving chess *problems*, and can easily outstrip humans at this endeavour.[6]

Chess-playing machines rely a lot on 'book knowledge' in addition to accurate calculational power. It is worth remarking that chess-playing machines fare better on the whole, relative to a comparable human player, when it is required that the moves are made very quickly; the human players perform relatively better in relation to the machines when a good measure of time is allowed for each move. One can understand this in terms of the fact that the computer's decisions are made on the basis of precise and rapid extended computations, whereas the human player takes advantage of 'judgements', that rely upon comparatively slow conscious assessments. These human judgements serve to cut down drastically the number of serious possibilities that need be considered at each stage of calculation, and much greater depth can be achieved in the analysis, when the time *is* available, than in the machine's simply calculating and directly eliminating possibilities, without using such judgements. (This difference is even more noticeable with the difficult Oriental game of 'go', where the number of possibilities per move is considerably greater than in chess.) The relationship between consciousness and the forming of judgements will be central to my later arguments, especially in Chapter 10.

AN AI APPROACH TO 'PLEASURE' AND 'PAIN'

One of the claims of AI is that it provides a route towards some sort of understanding of mental qualities, such as happiness, pain, hunger. Let us take the example of Grey Walter's tortoise. When its batteries ran low its behaviour pattern would change, and it would then act in a way designed to replenish its store of energy. There are clear analogies between this and the way that a human being – or any other animal – would act when feeling hungry. It perhaps might not be too much of a distortion of language to say that the Grey Walter tortoise was 'hungry' when it acted in this way. Some mechanism within it was sensitive to the state of charge in its battery, and when this got below a certain point it switched the tortoise over to a different behaviour pattern. No doubt there is something similar operating within animals when they become hungry, except that the changes in behaviour patterns are more complicated and subtle. Rather than simply switching over from one behaviour pattern to another, there is a change in *tendencies* to act in certain ways, these changes becoming stronger (up to a point) as the need to replenish the energy supply increases.

Likewise, some AI supporters envisage that concepts such as pain or happiness can be appropriately modelled in this way. Let us simplify things and consider just a single scale of 'feelings' ranging from extreme 'pain' (score: −100) to extreme 'pleasure' (score: +100). Imagine that we have a device – a machine of some kind, presumably electronic – that has a means of registering its own (putative) 'pleasure–pain' score, which I refer to as its 'pp-score'. The device is to have certain modes of behaviour and certain inputs, either internal (like the state of its batteries) or external. The idea is that its actions are geared so as to maximize its pp-score. There could be many factors which influence the pp-score. We could certainly arrange that the charge in its battery is one of them, so that a low charge counts negatively and a high charge positively, but there could be other factors too. Perhaps our device has some solar panels on it which give it an alternative means of obtaining energy, so that its batteries need not be used when the panels are in operation. We could arrange that by moving towards the light it can increase its pp-score a little, so that in the absence of other factors this is what it would tend to do. (Actually, Grey Walter's tortoise used to *avoid* the light!) It would need to have some means of performing computations so that it could work out the likely effects that different actions on its part would ultimately have on its pp-score. It could introduce probability weightings, so that a calculation would count as having a larger or smaller effect on the score depending upon the reliability of the data upon which it is based.

It would be necessary also to provide our device with other 'goals' than just maintaining its energy supply, since otherwise we should have no means of distinguishing 'pain' from 'hunger'. No doubt it is too much to ask that our device have a means of

procreation so, for the moment, sex is out! But perhaps we can implant in it a 'desire' for companionship with other such devices, by giving meetings with them a positive pp-score. Or we could make it 'crave' learning for its own sake, so that the mere storing of facts about the outside world would also score positively on its pp-scale. (More selfishly, we could arrange that performing various services for *us* have a positive score, as one would need to do if constructing a robot servant!) It might be argued that there is an artificiality about imposing such 'goals' on our device according to our whim. But this is not so very different from the way that natural selection has imposed upon us, as individuals, certain 'goals' which are to a large extent governed by the need to propagate our genes.

Suppose, now, that our device has been successfully constructed in accordance with all this. What right would we have to assert that it actually *feels* pleasure when its pp-score is positive and pain when the score is negative? The AI (or operational) point of view would be that we judge this simply from the way that the device behaves. Since it acts in a way which increases its score to as large a positive value as possible (and for as long as possible) and it correspondingly also acts to avoid negative scores, then we could reasonably *define* its feeling of pleasure as the degree of positivity of its score, and correspondingly *define* its feeling of pain to be the degree of negativity of the score. The 'reasonableness' of such a definition, it would be argued, comes from the fact that this is precisely the way that a human being reacts in relation to feelings of pleasure or pain. Of course, with human beings things are actually not nearly so simple as that, as we all know: sometimes we seem deliberately to court pain, or to go out of our way to avoid certain pleasures. It is clear that our actions are really guided by much more complex criteria than these (cf. Dennett 1978, pp. 190–229). But as a very rough approximation, avoiding pain and courting pleasure is indeed the way we act. To an operationalist this would be enough to provide justification, at a similar level of approximation, for the *identification* of pp-score in our device with its pain–pleasure rating. Such identifications seem also to be among the aims of AI theory.

We must ask: Is it really the case that our device would actually *feel* pain when its pp-score is negative and pleasure when it is positive? Indeed, could our device feel anything at all? The operationalist would, no doubt, either say 'Obviously yes', or dismiss such questions as meaningless. But it seems to me to be clear that there *is* a serious and difficult question to be considered here. In ourselves, the influences that drive us are of various kinds. Some are conscious, like pain or pleasure; but there are others of which we are not directly aware. This is clearly illustrated by the example of a person touching a hot stove. An involuntary action is set up which causes him to withdraw his hand even before he experiences any sensation of pain. It would seem to be the case that such involuntary actions are very much closer to the responses of our device to its pp-score than are the actual effects of pain or pleasure.

One often uses anthropomorphic terms in a descriptive, often jocular, way to describe the behaviour of machines: 'My car doesn't seem to want to start this morning'; or 'My watch still thinks it's running on Californian time'; or 'My computer claims it didn't understand that last instruction and doesn't know what to do next.' Of course we don't *really* mean to imply that the car actually might *want* something, or that the watch *thinks*, or that the computer* actually *claims* anything or that it *understands* or even *knows* what it is doing. Nevertheless such statements can be genuinely descriptive and helpful to our own understanding, provided that we take them merely in the spirit in which they are intended and do not regard them as literal assertions. I would take a rather similar attitude to various claims of AI that mental qualities might be present in the devices which have been constructed – *irrespective* of the spirit in which they are intended! If I agree to say that Grey Walter's tortoise can be hungry, it is in this half-jocular sense that I mean it. If I am prepared to use terms such as 'pain' or 'pleasure' for the pp-score of a device as envisaged above, it is because I find these terms helpful to my understanding of its behaviour, owing to certain analogies with my own behaviour and mental states. I do not mean to imply that these analogies are really particularly close or, indeed, that there are not

* As of 1989!

other *un*conscious things which influence my behaviour in a much *more* analogous way.

I hope it is clear to the reader that in my opinion there is a great deal more to the understanding of mental qualities than can be directly obtained from AI. Nevertheless, I do believe that AI presents a serious case which must be respected and reckoned with. In saying this I do not mean to imply that very much, if anything, has yet been achieved in the simulation of actual intelligence. But one has to bear in mind that the subject is very young. Computers will get faster, have larger rapid-access stores, more logical units, and will have large numbers of operations performed in parallel. There will be improvements in logical design and in programming technique. These machines, the vehicles of the AI philosophy, will be vastly improved in their technical capabilities. Moreover, the philosophy itself is *not* an intrinsically absurd one. Perhaps human intelligence can indeed be very accurately simulated by electronic computers – essentially the computers of today, based on principles that are already understood, but with the much greater capacity, speed, etc., that they are bound to have in the years to come. Perhaps, even, these devices will actually *be* intelligent; perhaps they will think, feel, and have minds. Or perhaps they will not, and some new principle is needed, which is at present thoroughly lacking. That is what is at issue, and it is a question that cannot be dismissed lightly. I shall try to present evidence, as best I see it. Eventually I shall put forward my own suggestions.

STRONG AI AND SEARLE'S CHINESE ROOM

There is a point of view, referred to as *strong AI* which adopts a rather extreme position on these issues.[7] According to strong AI, not only would the devices just referred to indeed be intelligent and have minds, etc., but mental qualities of a sort can be attributed to the logical functioning of *any* computational device, even the very simplest mechanical ones, such as a thermostat.[8] The idea is that mental activity is simply the carrying out of some well-defined sequence of operations, frequently referred to as an *algorithm*. I shall be more precise later on, as to what an algorithm actually is. For the moment, it will be adequate to define an algorithm simply as a calculational procedure of some kind. In the case of a thermostat, the algorithm is extremely simple: the device registers whether the temperature is greater or smaller than the setting, and then it arranges that the circuit be disconnected in the former case and connected in the latter. For any significant kind of mental activity of a human brain, the algorithm would have to be something vastly more complicated but, according to the strong-AI view, an algorithm nevertheless. It would differ very greatly in degree from the simple algorithm of the thermostat, but need not differ in principle. Thus, according to strong AI, the difference between the essential functioning of a human brain (including all its conscious manifestations) and that of a thermostat lies only in this much greater *complication* (or perhaps 'higher-order structure' or 'self-referential properties', or some other attribute that one might assign to an algorithm) in the case of a brain. Most importantly, all mental qualities – thinking, feeling, intelligence, understanding, consciousness – are to be regarded, according to this view, merely as aspects of this complicated functioning; that is to say, they are features merely of the *algorithm* being carried out by the brain.

The virtue of any specific algorithm would lie in its performance, namely in the accuracy of its results, its scope, its economy, and the speed with which it can be operated. An algorithm purporting to match what is presumed to be operating in a human brain would need to be a stupendous thing. But if an algorithm of this kind exists for the brain – and the supporters of strong AI would certainly claim that it does – then it could in principle be run on a computer. Indeed it could be run on *any* modern general-purpose electronic computer, were it not for limitations of storage space and speed of operation. (The justification of this remark will come later, when we come to consider the universal Turing machine.) It is anticipated that any such limitations would be overcome for the large fast computers of the not-too-distant future. In that eventuality, such an algorithm, if it could be found, would presumably pass the Turing test. The supporters of strong AI would claim that whenever the algorithm were run it

would, *in itself*: experience feelings; have a consciousness; be a mind.

By no means everyone would be in agreement that mental states and algorithms can be identified with one another in this kind of way. In particular, the American philosopher John Searle (1980, 1987) has strongly disputed that view. He has cited examples where simplified versions of the Turing test have actually *already* been passed by an appropriately programmed computer, but he gives strong arguments to support the view that the relevant mental attribute of 'understanding' is, nevertheless, entirely absent. One such example is based on a computer program designed by Roger Schank (Schank and Abelson 1977). The aim of the program is to provide a simulation of the understanding of simple stories like: 'A man went into a restaurant and ordered a hamburger. When the hamburger arrived it was burned to a crisp, and the man stormed out of the restaurant angrily, without paying the bill or leaving a tip.' For a second example: 'A man went into a restaurant and ordered a hamburger; when the hamburger came he was very pleased with it; and as he left the restaurant he gave the waitress a large tip before paying his bill.' As a test of 'understanding' of the stories, the computer is asked whether the man ate the hamburger in each case (a fact which had not been explicitly mentioned in either story). To this kind of simple story and simple question the computer can give answers which are essentially indistinguishable from the answers an English-speaking human being would give, namely, for these particular examples, 'no' in the first case and 'yes' in the second. So in this *very* limited sense a machine has already passed a Turing test!

The question that we must consider is whether this kind of success actually indicates any genuine understanding on the part of the computer – or, perhaps, on the part of the program itself. Searle's argument that it does *not* is to invoke his concept of a 'Chinese room'. He envisages first of all, that the stories are to be told in Chinese rather than English – surely an inessential change – and that all the operations of the computer's algorithm for this particular exercise are supplied (in English) as a set of instructions for manipulating counters with Chinese symbols on them. Searle imagines *himself* doing all the manipulations inside a locked room. The sequences of symbols representing the stories, and then the questions, are fed into the room through some small slot. No other information whatever is allowed in from the outside. Finally, when all the manipulations are complete, the resulting sequence is fed out again through the slot. Since all these manipulations are simply carrying out the algorithm of Schank's program, it must turn out that this final resulting sequence is simply the Chinese for 'yes' or 'no', as the case may be, giving the correct answer to the original question in Chinese about a story in Chinese. Now Searle makes it quite clear that he doesn't understand a word of Chinese, so he would not have the faintest idea what the stories are about. Nevertheless, by correctly carrying out the series of operations which constitute Schank's algorithm (the instructions for this algorithm having been given to him in English) he would be able to do as well as a Chinese person who would indeed understand the stories. Searle's point – and I think it is quite a powerful one – is that the mere carying out of a successful algorithm does *not* in itself imply that any understanding has taken place. The (imagined) Searle, locked in his Chinese room, would not understand a single word of any of the stories!

A number of objections have been raised against Searle's argument. I shall mention only those that I regard as being of serious significance. In the first place, there is perhaps something rather misleading in the phrase 'not understand a single word', as used above. Understanding has as much to do with patterns as with individual words. While carrying out algorithms of this kind, one might well begin to perceive something of the patterns that the symbols make without understanding the actual meanings of many of the individual symbols. For example, the Chinese character for 'hamburger' (if, indeed, there is such a thing) could be replaced by that for some other dish, say 'chow mein', and the stories would not be significantly affected. Nevertheless, it seems to me to be reasonable to suppose that in fact very little of the stories' actual meanings (even regarding such replacements as being unimportant) would come through if one merely kept following through the details of such an algorithm.

In the second place, one must take into account the fact that the execution of even a rather simple computer program would

normally be something extraordinarily lengthy and tedious if carried out by human beings manipulating symbols. (This is, after all, why we have computers to do such things for us!) If Searle were actually to perform Schank's algorithm in the way suggested, he would be likely to be involved with many days, months, or years of extremely boring work in order to answer just a single question – not an altogether plausible activity for a philosopher! However, this does not seem to me to be a serious objection since we are here concerned with matters of *principle* and not with practicalities. The difficulty arises more with a putative computer program which is supposed to have sufficient complication to match a human brain and thus to pass the Turing test *proper*. Any such program would have to be horrendously complicated. One can imagine that the operation of this program, in order to effect the reply to even some rather simple Turing-test question, might involve so many steps that there would be no possibility of any single human being carrying out the algorithm by hand within a normal human lifetime. Whether this would indeed be the case is hard to say, in the absence of such a program.[9] But, in any case, this question of extreme complication cannot, in my opinion, simply be ignored. It is true that we are concerned with matters of principle here, but it is not inconceivable to me that there might be some 'critical' amount of complication in an algorithm which it is necessary to achieve in order that the algorithm exhibit mental qualities. Perhaps this critical value is so large that no algorithm, complicated to that degree, could conceivably be carried out by hand by any human being, in the manner envisaged by Searle.

Searle himself has countered this last objection by allowing a whole team of human non-Chinese-speaking symbol manipulators to replace the previous single inhabitant ('himself') of his Chinese room. To get the numbers large enough, he even imagines replacing his room by the whole of India, its entire population (excluding those who understand Chinese!) being now engaged in symbol manipulation. Though this would be in practice absurd, it is not *in principle* absurd, and the argument is essentially the same as before: the symbol manipulators do *not* understand the story, despite the strong-AI claim that the mere carrying out of the appropriate algorithm would elicit the mental quality of 'understanding'. However, now another objection begins to loom large. Are not these individual Indians more like the individual neurons in a person's brain than like the whole brain itself? No-one would suggest that neurons, whose firings apparently constitute the physical activity of a brain in the act of thinking, would *themselves* individually understand what that person is thinking, so why expect the individual Indians to understand the Chinese stories? Searle replies to this suggestion by pointing out the apparent absurdity of India, the actual country, understanding a story that none of its individual inhabitants understands. A country, he argues, like a thermostat or an automobile, is not in the 'business of understanding', whereas an individual person is.

This argument has a good deal less force to it than the earlier one. I think that Searle's argument is at its strongest when there is just a single person carrying out the algorithm, where we restrict attention to the case of an algorithm which is sufficiently uncomplicated for a person actually to carry it out in less than a lifetime. I do *not* regard his argument as *rigorously* establishing that there is not some kind of disembodied 'understanding' associated with the person's carrying out of that algorithm, and whose presence does not impinge in any way upon his own consciousness. However, I would agree with Searle that this possibility has been rendered rather implausible, to say the least. I think that Searle's argument has a considerable force to it, even if it is not altogether conclusive. It is rather convincing in demonstrating that algorithms with the kind of complication that Schank's computer program possesses cannot have any genuine understanding whatsoever of the tasks that they perform; also, it *suggests* (but no more) that no algorithm, no matter how complicated, can ever, of itself alone, embody genuine understanding – in contradistinction to the claims of strong AI.

There are, as far as I can see, other very serious difficulties with the strong-AI point of view. According to strong AI, it is simply the algorithm that counts. It makes no difference whether that algorithm is being effected by a brain, an electronic computer, an entire country of Indians, a mechanical device of wheels and cogs, or a system of water pipes. The viewpoint is that it is simply the

logical structure of the algorithm that is significant for the 'mental state' it is supposed to represent, the particular physical embodiment of that algorithm being entirely irrelevant. As Searle points out, this actually entails a form of 'dualism'. *Dualism* is a philosophical viewpoint espoused by the highly influential seventeenth century philosopher and mathematician René Descartes, and it asserts that there are two separate kinds of substance: 'mind-stuff' and ordinary matter. Whether, or how, one of these kinds of substance might or might not be able to affect the other is an additional question. The point is that the mind-stuff is not supposed to be composed of matter, and is able to exist independently of it. The mind-stuff of strong AI is the logical structure of an algorithm. As I have just remarked, the particular physical embodiment of an algorithm is something totally irrelevant. The algorithm has some kind of disembodied 'existence' which is quite apart from any realization of that algorithm in physical terms. How seriously we must take this kind of existence is a question I shall need to return to in the next chapter. It is part of the general question of the Platonic reality of abstract mathematical objects. For the moment I shall sidestep this general issue and merely remark that the supporters of strong AI do indeed seem to be taking the reality at least of algorithms seriously, since they believe that algorithms form the 'substance' of their thoughts, their feelings, their understanding, their conscious perceptions. There is a remarkable irony in this fact that, as Searle has pointed out, the standpoint of strong AI seems to drive one into an extreme form of dualism, the very viewpoint with which the supporters of strong AI would least wish to be associated!

This dilemma lies behind the scenes of an argument put forward by Douglas Hofstadter (1981) – himself a major proponent of the strong-AI view – in a dialogue entitled 'A Conversation with Einstein's Brain'. Hofstadter envisages a book, of absurdly monstrous proportions, which is supposed to contain a complete description of the brain of Albert Einstein. Any question that one might care to put to Einstein can be answered, just as the living Einstein would have, simply by leafing through the book and carefully following all the detailed instructions it provides. Of course 'simply' is an utter misnomer, as Hofstadter is careful to point out. But his claim is that *in principle* the book is completely equivalent, in the operational sense of a Turing test, to a ridiculously slowed-down version of the actual Einstein. Thus, according to the contentions of strong AI, the book would think, feel, understand, be aware, just as though it were Einstein himself, but perhaps living at a monstrously slowed-down rate (so that to the book-Einstein the world outside would seem to flash by at a ridiculously speeded-up rate). Indeed, since the book is supposed to be merely a particular embodiment of the algorithm which constitutes Einstein's 'self', it would actually *be* Einstein.

But now a new difficulty presents itself. The book might never be opened, or it might be continually pored over by innumerable students and searchers after truth. How would the book 'know' the difference? Perhaps the book would not need to be opened, its information being retrieved by means of X-ray tomography, or some other technological wizardry. Would Einstein's awareness be enacted only when the book is being so examined? Would he be aware twice over if two people chose to ask the book the same question at two completely different times? Or would that entail two separate and temporally distinct instances of the *same* state of Einstein's awareness? Perhaps his awareness would be enacted only if the book is *changed*? After all, normally when we are aware of something we receive information from the outside world which affects our memories, and the states of our minds are indeed slightly changed. If so, does this mean that it is (suitable) *changes* in algorithms (and here I am including the memory store as part of the algorithm) which are to be associated with mental events rather than (or perhaps in addition to) the *activation* of algorithms? Or would the book-Einstein remain completely self-aware even if it were never examined or disturbed by anyone or anything? Hofstadter touches on some of these questions, but he does not really attempt to answer or to come to terms with most of them.

What does it mean to activate an algorithm, or to embody it in physical form? Would changing an algorithm be different in any sense from merely discarding one algorithm and replacing it with another? What on earth does any of this have to do with our feelings of conscious awareness? The reader (unless himself or

herself a supporter of strong AI) may be wondering why I have devoted so much space to such a patently absurd idea. In fact, I do *not* regard the idea as intrinsically an absurd one – mainly just wrong! There is, indeed some force in the reasoning behind strong AI which must be reckoned with, and this I shall try to explain. There is, also, in my opinion, a certain appeal in some of the ideas – if modified appropriately – as I shall also try to convey. Moreover, in my opinion, the particular contrary view expressed by Searle also contains some serious puzzles and seeming absurdities, even though, to a partial extent, I agree with him!

Searle, in his discussion, seems to be implicitly accepting that electronic computers of the present-day type, but with considerably enhanced speed of action and size of rapid-access store (and possibly parallel action) may well be able to pass the Turing test proper, in the not-too-distant future. He is prepared to accept the contention of strong AI (and of most other 'scientific' viewpoints) that 'we are the instantiations of any number of computer programs'. Moreover, he succumbs to: 'Of course the brain is a digital computer. Since everything is a digital computer, brains are too.'[10] Searle maintains that the distinction between the function of human brains (which can have minds) and of electronic computers (which, he has argued, cannot) both of which might be executing the same algorithm, lies solely in the material construction of each. He claims, but for reasons he is not able to explain, that the biological objects (brains) can have 'intentionality' and 'semantics', which he regards as defining characteristics of mental activity, whereas the electronic ones cannot. In itself this does not seem to me to point the way towards any helpful scientific theory of mind. What is so special about biological systems, apart perhaps from the 'historical' way in which they have evolved (and the fact that *we* happen to be such systems), which sets them apart as the objects allowed to achieve intentionality or semantics? The claim looks to me suspiciously like a dogmatic assertion, perhaps no less dogmatic, even, than those assertions of strong AI which maintain that the mere enacting of an algorithm can conjure up a state of conscious awareness!

In my opinion Searle, and a great many other people, have been led astray by the computer people. And they, in turn, have been led astray by the physicists. (It is not the physicists' fault. Even *they* don't know everything!) The belief seems to be widespread that, indeed, 'everything is a digital computer'. It is my intention, in this book, to try to show why, and perhaps how, this need *not* be the case.

HARDWARE AND SOFTWARE

In the jargon of computer science, the term *hardware* is used to denote the actual machinery involved in a computer (printed circuits, transistors, wires, magnetic storage space, etc.), including the complete specification for the way in which everything is connected up. Correspondingly, the term *software* refers to the various programs which can be run on the machine. It was one of Alan Turing's remarkable discoveries that, in effect, any machine for which the hardware has achieved a certain definite degree of complication and flexibility, is *equivalent* to any other such machine. This equivalence is to be taken in the sense that for any two such machines A and B there would be a specific piece of software which if given to machine A would make it act precisely as though it were machine B; likewise, there would be another piece of software which would make machine B act precisely like machine A. I am using the word 'precisely' here to refer to the actual output of the machines for any given input (fed in after the converting software is fed in) and *not* to the *time* that each machine might take to produce that output. I am also allowing that if either machine at any stage runs out of storage space for its calculations then it can call upon some (in principle unlimited) external supply of blank 'rough paper' – which could take the form of magnetic tape, discs, drums or whatever. In fact, the difference in the time taken by machines A and B to perform some task, might well be a very serious consideration. It might be the case, for example, that A is more than a thousand times faster at performing a particular task than B. It might also be the case that, for the very same machines, there is some other task for which B is a thousand times faster than A. Moreover, these timings could depend very greatly on the particular choices of converting software that are used. This is very much an 'in-principle' discussion, where one is not really concerned with such practical matters as achieving one's calculations in a reasonable time. I shall be more precise in the next section about the concepts being referred to here: the machines A and B are instances of what are called *universal Turing machines*.

In effect, all modern general purpose computers are universal Turing machines. Thus, all general purpose computers are equivalent to one another in the above sense: the differences between them can be entirely subsumed in the software, provided that we are not concerned about differences in the resulting speed of operation and possible limitations on storage size. Indeed, modern technology has enabled computers to perform so swiftly and with such vast storage capacities that, for most 'everyday' purposes, neither of these practical considerations actually represents any serious limitation to what is normally needed,* so this effective theoretical equivalence between computers can also be seen at the practical level. Technology has, it seems, transformed entirely academic discussions concerning idealized computing devices into matters which directly affect all our lives!

As far as I can make out, one of the most important factors underlying the strong-AI philosophy is this equivalence between physical computing devices. The hardware is seen as being relatively unimportant (perhaps even totally unimportant) and the software, i.e. the program, or the algorithm, is taken to be the one vital ingredient. However, it seems to me that there are also other important underlying factors, coming more from the direction of physics. I shall try to give some indication of what these factors are.

What is it that gives a particular person his individual identity? Is it, to some extent, the very atoms that compose his body? Is his identity dependent upon the particular choice of electrons, protons, and other particles that compose those atoms? There are at least two reasons why this cannot be so. In the first place, there is a continual turnover in the material of any living person's body. This applies in particular to the cells in a person's brain, despite

* However, see the discussion of complexity theory and NP problems at the end of Chapter 4.

the fact that no new actual brain cells are produced after birth. The vast majority of atoms in each living cell (including each brain cell) – and, indeed, virtually the entire material of our bodies – has been replaced many times since birth.

The second reason comes from quantum physics – and by a strange irony is, strictly speaking, in contradiction with the first! According to quantum mechanics (and we shall see more about this in Chapter 6, p. 360), any two electrons must necessarily be completely identical, and the same holds for any two protons and for any two particles whatever, of any one particular kind. This is not merely to say that there is no way of telling the particles apart: the statement is considerably stronger than that. If an electron in a person's brain were to be exchanged with an electron in a brick, then the state of the system would be *exactly*[11] *the same state* as it was before, not merely indistinguishable from it! The same holds for protons and for any other kind of particle, and for whole atoms, molecules, etc. If the entire material content of a person were to be exchanged with corresponding particles in the bricks of his house then, in a strong sense, nothing would have happened whatsoever. What distinguishes the person from his house is the *pattern* of how his constituents are arranged, not the individuality of the constituents themselves.

There is perhaps an analogue of this at an everyday level, which is independent of quantum mechanics, but made particularly manifest to me as I write this, by the electronic technology which enables me to type at a word-processor. If I desire to change a word, say to transform 'make' into 'made', I may do this by simply replacing the 'k' by a 'd', or I may choose instead to type out the whole word again. If I do the latter, is the 'm' the same 'm' as was there before, or have I replaced it with an identical one? What about the 'e'? Even if I do simply replace 'k' by 'd', rather than retype the word, there is a moment just between the disappearance of 'k' and appearance of 'd' when the gap closes and there is (or, at least, sometimes is) a wave of re-alignment down the page as the placement of every succeeding letter (including the 'e') is re-calculated, and then re-re-calculated as the 'd' is inserted. (Oh, the cheapness of mindless calculation in this modern age!) In any case, *all* the letters that I see before me on the screen are mere gaps

in the track of an electron beam as the whole screen is scanned sixty times each second. If I take any letter whatever and replace it by an identical one, is the situation the *same* after the replacement, or merely indistinguishable from it? To try to adopt the second viewpoint (i.e. 'merely indistinguishable') as being distinct from the first (i.e. 'the same') seems footling. At least, it seems reasonable to call the situation the same when the letters are the same. And so it is with the quantum mechanics of identical particles. To replace one particle by an identical one is actually to have done nothing to the state at all. The situation is indeed to be regarded as the *same* as before. (However, as we shall see in Chapter 6, the distinction is actually *not* a trivial one in a quantum-mechanical context.)

The remarks above concerning the continual turnover of atoms in a person's body were made in the context of classical rather than quantum physics. The remarks were worded as though it might be meaningful to maintain the individuality of each atom. In fact classical physics is adequate and we do not go badly wrong, at this level of description, by regarding atoms as individual objects. Provided that the atoms are reasonably well separated from their identical counterparts as they move about, one *can* consistently refer to them as maintaining their individual identities since each atom can be, in effect, tracked continuously, so that one could envisage keeping a tab on each separately. From the point of view of quantum mechanics it would only be a convenience of speech to refer to the individuality of the atoms, but it is a consistent enough description at the level just considered.

Let us accept that a person's individuality has nothing to do with any individuality that one might try to assign to his material constituents. Instead, it must have to do with the *configuration*, in some sense, of those constituents – let us say the configuration in space or in space–time. (More about that later.) But the supporters of strong AI go further than this. If the information content of such a configuration can be translated into another form from which the original can again be recovered then, so they would claim, the person's individuality must remain intact. It is like the sequences of letters I have just typed and now see displayed on the screen of my word-processor. If I move them off the screen, they remain coded in the form of certain tiny displacements of electric charge, in some configuration in no clear way geometrically resembling the letters I have just typed. Yet, at any time I can move them back on to the screen, and there they are, just as though no transformation had taken place. If I choose to save what I have just written, then I can transfer the information of the sequences of letters into configurations of magnetization on a disc which I can then remove, and then by switching off the machine I neutralize all the (relevant) tiny charge displacements in it. Tomorrow, I can re-insert the disc, reinstate the little charge displacements and display the letter sequences again on the screen, just as though nothing had happened. To the supporters of strong AI, it is 'clear' that a person's individuality can be treated in just the same way. Like the sequences of letters on my display screen, so these people would claim, nothing is lost of a person's individuality – indeed nothing would really have happened to it at all – if his physical form were to be translated into something quite different, say into fields of magnetization in a block of iron. They appear even to claim that the person's conscious awareness would persist while the person's 'information' is in this other form. On this view, a 'person's awareness' is to be taken, in effect, as a piece of software, and his particular manifestation as a material human being is to be taken as the operation of this software by the hardware of his brain and body.

It seems that the reason for these claims is that, whatever material form the hardware takes – for example some electronic device – one could always 'ask' the software questions (in the manner of a Turing test), and assuming that the hardware performs satisfactorily in computing the replies to these questions, these replies would be identical to those that the person would make whilst in his normal state. ('How are you feeling this morning?' 'Oh, fairly well, thank you, though I have a slightly bothersome headache.' 'You don't feel, then, that there's . . . er . . . anything odd about your personal identity . . . or something?' 'No; why do you say that? It seems rather a strange question to be asking.' 'Then you feel yourself to be the same person that you were yesterday?' 'Of course I do!')

An idea frequently discussed in this kind of context is the

teleportation machine of science fiction.[12] It is intended as a means of 'transportation' from, say, one planet to another, but whether it actually would be such, is what the discussion is all about. Instead of being physically transported by a spaceship in the 'normal' way, the would-be traveller is scanned from head to toe, the accurate location and complete specification of every atom and every electron in his body being recorded in full detail. All this information is then beamed (at the speed of light), by an electromagnetic signal, to the distant planet of intended destination. There, the information is collected and used as the instructions to assemble a precise duplicate of the traveller, together with all his memories, his intentions, his hopes, and his deepest feelings. At least that is what is expected; for every detail of the state of his brain has been faithfully recorded, transmitted, and reconstructed. Assuming that the mechanism has worked, the original copy of the traveller can be 'safely' destroyed. Of course the question is: is this *really* a method of travelling from one place to another or is it merely the construction of a duplicate, together with the murder of the original? Would *you* be prepared to use this method of 'travel' – assuming that the method had been shown to be completely reliable, within its terms of reference? If teleportation is *not* travelling, then what is the difference *in principle* between it and just walking from one room into another? In the latter case, are not one's atoms of one moment simply providing the information for the locations of the atoms of the next moment? We have seen, after all, that there is no significance in preserving the identity of any particular atom. The question of the identity of any particular atom is not even meaningful. Does not any moving pattern of atoms simply constitute a kind of wave of information propagating from one place to another? Where is the essential difference between the propagation of waves which describes our traveller ambling in a commonplace way from one room to the other and that which takes place in the teleportation device?

Suppose it is true that teleportation *does* actually 'work', in the sense that the traveller's own 'awareness' is actually reawakened in the copy of himself on the distant planet (assuming that this question has genuine meaning). What would happen if the *original* copy of the traveller were not destroyed, as the rules of this game demand? Would his 'awareness' be in two places at once? (Try to imagine your response to being told the following: 'Oh dear, so the drug we gave you before placing you in the Teleporter has worn off prematurely has it? That is a little unfortunate, but no matter. Anyway, you will be pleased to hear that the other you – er, I mean the *actual* you, that is – has now arrived safely on Venus, so we can, er, dispose of you here – er, I mean of the *redundant* copy here. It will, of course, be quite painless.') The situation has an air of paradox about it. Is there anything in the laws of physics which could render teleportation *in principle* impossible? Perhaps, on the other hand, there is nothing in principle against transmitting a person, and a person's consciousness, by such means, but that the 'copying' process involved would inevitably destroy the original? Might it then be that the preserving of *two* viable copies is what is impossible in principle? I believe that despite the outlandish nature of these considerations, there *is* perhaps something of significance concerning the physical nature of consciousness and individuality to be gained from them. I believe that they provide one pointer, indicating a certain essential role for *quantum mechanics* in the understanding of mental phenomena. But I am leaping ahead. It will be necessary to return to these matters after we have examined the structure of quantum theory in Chapter 6 (cf. p. 348).

Let us see how the point of view of strong AI relates to the teleportation question. We shall suppose that somewhere between the two planets is a relay station, where the information is temporarily stored before being re-transmitted to its final destination. For convenience, this information is not stored in human form, but in some magnetic or electronic device. Would the traveller's 'awareness' be present in association with this device? The supporters of strong AI would have us believe that this must be so. After all, they say, any question that we might choose to put to the traveller could in principle be answered by the device, by 'merely' having a simulation set up for the appropriate activity of his brain. The device would contain all the necessary information; and the rest would just be a matter of computation. Since the device would reply to questions exactly as though it were the

traveller, then (Turing test!) it would *be* the traveller. This all comes back to the strong-AI contention that the actual hardware is not important with regard to mental phenomena. This contention seems to me to be unjustified. It is based on the presumption that the brain (or the mind) is, indeed, a digital computer. It assumes that no specific physical phenomena are being called upon, when one thinks, that might demand the particular physical (biological, chemical) structure that brains actually have.

No doubt it would be argued (from the strong-AI point of view) that the only assumption that is really being made is that the effects of any specific physical phenomena which need to be called upon can always be accurately *modelled* by digital calculations. I feel fairly sure that most physicists would argue that such an assumption is actually a very natural one to make on the basis of our present physical understandings. I shall be presenting the reasons for my own contrary view in later chapters (where I shall also need to lead up to why I believe that there is even any appreciable assumption being made). But, just for the moment, let us accept this (commonly held) view that all the relevant physics *can* always be modelled by digital calculations. Then the only real assumption (apart from questions of time and calculation space) is the 'operational' one that if something *acts* entirely like a consciously aware entity, then one must also maintain that it 'feels' itself to be that entity.

The strong-AI view holds that, being 'just' a hardware question, any physics actually being called upon in the workings of the brain can necessarily be simulated by the introduction of appropriate converting software. If we accept the operational viewpoint, then the question rests on the equivalence of universal Turing machines, and on the fact that any algorithm can, indeed, be effected by such a machine – together with the presumption that the brain acts according to some kind of algorithmic action. It is time for me to be more explicit about these intriguing and important concepts.

NOTES

1. See, for example, Gardner (1958), Gregory (1981), and references contained therein.
2. See, for example, Resnikoff and Wells (1984), pp. 181–4. For a classic account of calculating prodigies generally, see Rouse Ball (1892); also Smith (1983).
3. See Gregory (1981), pp. 285–7, Grey Walter (1953).
4. This example is quoted from Delbrück (1986).
5. See the articles by O'Connell (1988) and Keene (1988). For more information about computer chess, see Levy (1984).
6. Of course, most chess problems are designed to be hard for *humans* to solve. It would probably not be too difficult to construct a chess problem that human beings would not find enormously hard, but which present-day chess-problem solving computers could not solve in a thousand years. (What would be required would be a fairly obvious plan, a very large number of moves deep. Problems are known, for example, requiring some 200 moves – more than enough!) This suggests an interesting challenge.
7. Throughout this book I have adopted Searle's terminology 'strong AI' for this extreme viewpoint, just to be specific. The term 'functionalism' is frequently used for what is essentially the same viewpoint, but perhaps not always so specifically. Some proponents of this kind of view are Minsky (1968), Fodor (1983), Hofstadter (1979), and Moravec (1989).
8. See Searle (1987), p. 211, for an example of such a claim.
9. In his criticism of Searle's original paper, as reprinted in *The Mind's I*, Douglas Hofstadter complains that no human being could conceivably 'internalize' the entire description of another human being's mind, owing to the complication involved. Indeed not! But as I see it, that is not entirely the point. One is concerned merely with the carrying out of that part of an algorithm which purports to embody the occurrence of a single mental event. This could be some momentary 'conscious realization' in the answering of a Turing-test question, or it could be something simpler. Would any such 'event' necessarily require an algorithm of stupendous complication?
10. See pp. 368, 372 in Searle's (1980) article in Hofstadter and Dennett (1981).
11. Some readers, knowledgeable about such matters, might worry about a certain sign difference. But even that (arguable) distinction disappears if we rotate one of the electrons completely through

360° as we make the interchange! (See Chapter 6, p. 360 for an explanation.)

12. See the Introduction to Hofstadter and Dennett (1981).