1 Introduction

What! Out of senseless Nothing to provoke,
A conscious Something to resent the Yoke!
Edward Fitzgerald (*Omar Khayyam*)

'The sum of all the parts of Such –
Of each laboratory scene –
Is such.' While science means this much
And means no more, why, let it mean!
But were the science-men to find
Some animating principle
Which gave synthetic Such a mind
Vital, though metaphysical –
To Such, such an event, I think,
Would cause unscientific pain:
Science, appalled by thought, would shrink
To its component parts again.
Robert Graves (*Synthetic Such*)

A forbearing machine

In recent months, I have been periodically guilty of what would normally be looked upon as bad behaviour, namely the making of audible comments during games of chess. Seated across the board from my commendably silent opponent, I have been voicing remarks about my adversary's moves, and indulging in outbursts of self-approbation whenever it appeared that I was gaining the upper hand.

Had these games been played in a chess club, I would no doubt have been asked to leave the premises, and even though they actually took place in my own home, the conduct nevertheless seems reprehensible. So why have my taunts been borne with such forbearance? Why, in the face of such provocation, has my rival continued to display mute equanimity? For the simple reason that it is a machine: a computer, which happens to have a chess program in its memory.

It is a sign of the times that I must qualify these statements, by adding that my computer is equipped with neither a speech synthesizer nor a contraption for receiving and analysing the human voice. Both types of device are now conceivable attachments to such machines. But even though it is thus not out of the question for a chess-playing computer to be able to give as good as it receives, verbally, the fact remains that it would be unlikely to react in the way that would be quite normal for a human being; it would probably not refuse to play with a partner prone to such boorish manners.

But a machine could be built which would *physically* respond in just

that fashion. Robots exist which are capable of walking, and although this represents a task that is far from trivial, it would be possible, in principle, to furnish such a device with a means of receiving my words, of evaluating their meaning, of perceiving my undesirability as a chess opponent, and of simply leaving the board and walking away. In so doing, it would be reacting in what could broadly be called a human manner. But most people would shy from the idea of calling it a pseudo-person; they would be unlikely to accord the robot human status despite anything that was done to make it more sophisticated.

What is the origin of our scepticism, when confronted with this issue? Why are we so reluctant to allow the possibility that a machine could think, in much the same way that we do? What is the quality that is so quintessentially human that it defies replication in any man-made machine, however complex? I believe that most people would opt for *consciousness*, (or possibly *self*-consciousness[a]). This is the capacity that seems to belong so uniquely to *Homo sapiens*, and there are some who are loathe to credit even the other animals with it, let alone machines. There are others, however, who feel that it will one day be possible to build conscious devices, and some have already striven to give computers what is generally referred to as *artificial intelligence*.

Is that a realistic quest, or is it doomed to fail, because of an inability to grasp some fundamental difference between the human mind and anything that could be fabricated by technology? This will be one of the themes of the present book, and I will be seeking to put the reader in a better position to judge the issue. To judge it, that is, not from the standpoint of traditional artificial intelligence (for that venerable field of technology has already been the subject of considerable discussion), but rather through an appreciation of what are known as *neural networks*.[b] And part of my job will be to explain just what is meant by that expression, indeed, for it would be easy to be confused by the coupling of a biological word with another that so smacks of technology.

Let me state immediately, therefore, that the term neural network is commonly applied in a variety of contexts. These span a spectrum that stretches from the neural circuitry we have in our brains to computers in which some of the components are wired up in a manner reminiscent of the way the brain's nerve cells are interconnected. In particular, and when it is preceded by the qualifier *artificial*, the term is used to describe certain computational strategies that are inspired by the brain's structure. Such strategies have scored a number of successes in recent years, and I will describe some particularly prominent examples. The discussion of artificial neural networks will also serve to illustrate certain principles that these versions have bor-

a / Self-consciousness in this context refers to the ability to examine one's thoughts, rather than to shyness.
b / A commonly encountered alternative term is *parallel distributed systems*. The neural enterprise, both in real brains and as a computational strategy, is broadly referred to by the terms *connectionism*, *computational neuroscience* and *neural computation*.

rowed from the brain itself. It must be emphasized, however, that there are fundamental differences between neural computer hardware and computational strategies, on the one hand, and the genuine article on the other. I will be drawing a careful distinction between the real and the artificial in this book, but I will also be venturing to judge whether the dividing line is ever likely to be breached.

Research on the real brain has of course being going on for centuries. There are reports on the effects of damage to that organ which date from the second millennium BC. Attempts at theoretically modelling the brain are on the point of being able to celebrate their hundredth jubilee, whereas computational strategies employing artificial neural networks arrived on the scientific stage only a few decades ago. Brain-inspired computer hardware is of even more recent vintage; at the time this book is being written, the marketing of such devices is still in its infancy.

The present book's title was taken from a passage in Charles Sherrington's *Man on his Nature*. Sherrington's contributions to the study of the brain were made during a remarkably productive scientific career which stretched over almost seventy years. The revised edition of his seminal book was published in 1952, a year before his death.[1] At that time, the systematic study of the brain's neural networks was barely under way, and it is not surprising that even a man of Sherrington's intellectual stature should have looked upon the brain's workings as something almost magical. But the view of the brain to which he and his contemporaries had been contributing allowed precious little room for the mysterious; with each new advance, the picture was becoming increasingly mechanistic. One can imagine Sherrington and his colleagues speculating about the possibility of constructing a machine capable of thought.

The idea had already been in the air for some time, even in his day. Primitive robots had been built, and there was also the stimulation from literature. A famous step along that road was taken in the story by Mary Wollstonecraft Shelley, published in 1818, in which a certain Dr Frankenstein fabricates a creature from parts of dead humans, and galvanizes the completed body to life by subjecting it to a jolt of electricity. That novel gave birth to an entirely new genre, and the idea that intelligent machines can be created is now so common in fiction that it almost leaves us blasé.[c]

Because we nevertheless draw a sharp distinction between fact and fantasy, confrontation with an actual robot, particularly one of the advanced type, tends to intrigue us. I recall seeing examples at the international fair Expo '85, held at Tsukuba, near Tokyo. Because the organizers wished to give those attending a taste of things to come, this exhibition was well stocked with all manner of clever technological achievements, but the real crowd

c / Another human trait that has recently been observed in clever machines is their susceptibility to 'infection', as typified by the notorious viruses that now plague so many home and institutional computers.

Figure 1.1
This sketch of the author was made by a portrait-drawing robot constructed by a research team at the National Panasonic Division of the Matsushita Electrical Industrial Company. The robot scanned the scene with a television camera, and its image processing unit evaluated the variations between the lighted and shaded parts of the original image. It then extracted the contours in order to form the elements of the line diagram. The associated image processing was carried out by a 16-bit microcomputer in about ten seconds. The early part of the human visual system is known to carry out a similar extraction of line elements from the visual scene, and it has also been shown that there are neurons deeper in the visual pathway which respond to juxtapositions of lines. As has now been established experimentally, there are neurons lying even deeper in the primate brain which exclusively react to a highly specific object, such as a familiar face.

attractors were the robots. And it was easy to locate them, because long queues formed outside the pavilions where they were being put through their paces. At one of these, I had the eerie experience of sitting in front of a portrait-drawing robot constructed by a research team at the National Panasonic Division of the Matsushita Electrical Industrial Company.

Once I had settled in the chair, it moved toward me and gazed at me with a video camera reminiscent of a Cyclopean eye. Then, after a brief pause, which in itself was a little unnerving because one wondered what it was up to, the machine started to make a drawing of me. But the impressive thing was that this was an outline sketch, with all the highlights and shadings normally present in a photographic image removed (see figure 1.1). This is no mean accomplishment for a machine. The image processing unit has to gauge the variations between the lighted and shaded parts of the original image, and it must then extract the contours in order to form the elements of the line diagram. (As will be discussed in a later chapter, it is believed that similar extraction occurs during the early part of the processing of visual information by our own brains.) In this artistic robot, the job of processing the information was carried out by a 16-bit microcomputer in about ten seconds.[d]

Another of the exhibits was still more impressive, in retrospect at least, even though the robot in question looked singularly cumbersome. The ungainly impression stemmed from the fact that this device was attempting to do something that seemed rather commonplace, namely mount a staircase. On closer inspection, however, one began to realize why it held the crowd so enthralled. It turned out that the staircase had been deliberately constructed with uneven step heights, and the machine would soon have toppled over if it had not been able to allow for this lack of uniformity. It too was a one-eyed monster, and the thing that most fascinated the large group of spectators was the way it would pause and seem to cogitate each time it had inspected a new step.

As impressed as we undoubtedly are by such feats of technical élan, we reserve our strongest reactions for those cases in which there is even the suggestion of a personality within the machine. Mere mechanical prowess, however complex, pales in comparison with the implication that such a device even has a mind of its own. This idea too is quite common in literature, of course, a striking example being that of the computer called HAL in Arthur C. Clarke's novel *2001: A Space Odyssey*.[e] Its job was ostensibly to control such things as the navigation and life-support systems in a space ship, and it was equipped with a speech synthesizer, to ease communication with the crew. But although this enabled it to exchange English words with the humans on board, that in itself was not particularly unusual; as noted earlier, a speech capability is now an optional extra for some personal computers. What made

d / The most important pieces of computer jargon are defined and explained in the glossary, as are the key biological terms.
e / Other prominent examples include the robots *R2D2* and *C3PO* in the *Star Wars* series.

HAL different was that it began to make decisions based on what could be called its own motives. And where these were in conflict with those of the crew, it gave priority to its own interests.

We have reached a key issue, for what could be more characteristic of a thinking machine than that it use its thoughts to further its own aims.[f] Is it not this, above all, that is the hallmark of the intelligent agency? But just how are a machine's aims to be defined, and how would one judge the intelligence of something which nevertheless has different prerequisites than that of a human being? These are just some of the points that will have to be discussed in this book.

Intelligence and consciousness – a first look

The concept of *intelligence* is so familiar that we feel we know exactly what is meant by the word. When pressed for a definition, however, we are likely to discover that it is easier to be aware of intelligence than to say just what it is. Our surprise at being thus stumped is likely to be compounded by a frustration stemming from the knowledge that some people not only recognize intelligence in their fellow humans but even manage to quantify it. The existence of intelligence tests, and the general acceptance of the idea of an *IQ*, tell us that this is the case.[2] Why then does delineation of the concept cause us so much trouble?

We need feel no embarrassment at this perceived inadequacy, for the fact is that even the professional measurers of intelligence do not agree on just what constitutes this obviously desirable characteristic. The manner in which a person's IQ should be gauged is in fact the subject of no small amount of dispute. There are many, indeed, who feel that the use of a single number to denote such a complicated concept is ludicrous. The more cautious of those active in this difficult branch of science identify several different facets of intelligence, and maintain that it must therefore be looked upon as a composite endowment.

With an anxious eye on the machines that may one day usurp us, we might find it easier to say what intelligence is *not*. The insatiable appetite that the typical computer displays for handling large amounts of data, and our feeling that these machines pose no immediate threat, suggests that mere manipulation of numbers and symbols cannot be accepted as a criterion of intelligence. We would be looking for some evidence of what could be called *wisdom*. This, surely, is what puts even the less-educated rustic well above anything that can be presently produced by the computer manufacturer. What is it that gives a person common sense, however, if it is not the ability to take note of the relationships between cause and effect, and act accordingly? And are these not reducible to the very manipulations of symbols that we might be tempted to denigrate when we see them performed by a computer?

f / The word *aims* should not be interpreted in a narrow, egoistic sense; I intend it to include altruistic goals.

Perhaps the most telling aspect of what we call intelligence is the capacity for reacting to novel situations. But the suitability of an individual's response would always have to be judged against the yardstick of what has gone before, even though the conditions that have to be coped with might be unique and thus without precedent. In short, the defining features of intelligence are the faculty for learning from experience, and the ability to apply acquired knowledge to fresh circumstances.

This brief definition circumscribes the concept, but it is insufficiently detailed to serve as a means of ascertaining the presence or absence of intelligence. Because such detection lies at the heart of what this book attempts to accomplish, we will have to delve deeper into the ramifications of sagacity and construct a set of practical rules by which it can be recognized. This is going to be forced upon us by the ultimate need to say whether or not thinking machines will ever be fabricated, but the demand is not a new one. On the contrary, it is the same requirement that burdens those who attempt to measure intelligence in our species. And the fact that such measurements aim at finding a specific level amongst a continuous spectrum of levels endorses the view that intelligence can only be adequately evaluated by taking many factors into account simultaneously. Intelligence, in other words, is not an all-or-nothing thing; each of us possesses intelligence to a certain degree. Let us take a closer look at some of its constituent factors, and let us use the word *agency* to cover both human and machine.

As we have already seen, the intelligent agency must be able to learn from experience, and it must be able to use that experience in making decisions. And the latter will, if they are to be more than mere caprice, involve foreseeing the possible repercussions of given courses of action. Such anticipation is impressive, because it unavoidably entails the ability both to imagine a variety of scenarios and to discern between consequences that are similar but not identical. The handling of novel situations thus implies categorization, association and generalization.

Turning to the question of how intelligence is actually evaluated, we find a situation that is far from ideal. For there are a variety of measures, each reflecting the emphases of their advocates. Indeed, they have sometimes also reflected the interests and exigencies that raised the need for intelligence testing in the first place. A case in point occurred in Paris in 1905, when the school authorities asked Alfred Binet to supply them with criteria for weeding out children with intelligence levels so low that they were unlikely to benefit from formal education.[3] To his credit, Binet devised a series of tests, thirty in all, which were unbiased by any education that the candidate had already received. And he rationalized his observations by defining a person's *mental age* as that of the majority of subjects who displayed the same level of ability as that of the one in question. He went on to recommend use of an intelligence quotient, IQ for short, which is the ratio of the mental age to the *biological age*, expressed as a percentage. Precocious children, well in advance of their years, thus revealed themselves as having IQs much greater than 100, while their poorly equipped counterparts had IQs considerably below the hundred mark.

Even this brief formal definition exposes a limitation in the context of our discussion, because it could not be applied to a machine. What, for example, are we to use for the biological age of a computer? The time that has elapsed since it emerged from the production line? The quite defensible use of biological age when determining the IQ of a human reminds us of the important fact that organisms evolve. But where is the counterpart of this process in a machine? In the case of a computer that possesses a program modifiable by experience, it might be justifiable to equate biological age with the number of opportunities that the program has been given of changing itself, but machine intelligence would have to be judged in a more general manner.

This was a problem that intrigued Alan Turing in the late 1940s, and he came up with a novel solution which he called the imitation game.[4] It can be couched in other terms, but let us consider it in the version originally put forward by Turing himself. The game is played by three individuals: a woman, a man, and an interrogator. The latter may be of either sex, and he or she is located in a room separate from that occupied by the other two. All communication proceeds exclusively by a teleprinter, to avoid the divulging of identity by tone of voice. (Had he been alive today, Turing would probably have replaced the latter by interlinked computer terminals.) The point of the game is that the man must attempt to hoodwink the interrogator into believing that he is a woman, while the real woman merely responds in such a way as to not reveal her gender explicitly. To that end, the two of them are even permitted to lie. Turing believed that a machine can be credited with intelligence if, having replaced the man in this game, it does just as well in fooling the interrogator as to its identity. The candidate machine would obviously have to be rather sophisticated, because it would need to be able to handle language well enough not to give the game away.[g]

Like several other of Turing's inventions, the machine in this game was merely part of an intellectual exercise; he did not actually build such a device. And there is an aspect of this thought experiment that is not well-defined, namely the amount of time that should be allowed to elapse before an assessment is made of the machine's performance; prolonged indefinitely, the test would almost certainly lead to the machine being unmasked. The point that Turing was trying to make was that verbal interaction with a real human serves as a reliable indicator of intelligence if it can be maintained for a reasonable period of time. We should note, too, that the ground rules in this game were not chosen arbitrarily. Turing could have settled on any of a wide variety of attributes, against which the machine's performance was to be appraised, but he settled on something that has almost endless ramifications, namely the difference between the human sexes. The myriad facets of this

g / I have heard remarks to the effect that it was fortunate that Turing did not ask the masquerader to feign being a psychiatrist, since every question would then have been convincingly parried with a *but what do YOU think*.

difference, which permeate the very fabric of human culture, are all potential pitfalls for the aspiring machine.

In contrast, any test that has well-defined rules, however intricate, will readily fall within the capabilities of a sufficiently sophisticated device, so such a task will not be able to expose the machine for what it is. This is an important point, in the present context, because the critical step in training an artificial neural network is invariably the selection of appropriate representations for the input and the output. And yet the discovery of suitable representations is tantamount to exposing the fact that the things being symbolized indeed *have* their rationale. When we later discuss examples of artificial neural networks in action, therefore, we will have to appreciate that the very tractability of a given training assignment might undermine the claim that the network functions intelligently. Turing knew perfectly well what he was doing when he chose the male–female schism as the basis for his experiment. How, for example, would one go about training an artificial network to emulate the proverbial woman's wiles?

No neural network has yet been able to match human behaviour for a sustained period. It seems prudent, therefore, to settle for the lesser aim of demonstrating that current artificial neural networks possess some rudimentary *aspects* of intelligence, such as learning from experience, even though they are not capable of displaying all the characteristics of the human variety. We will see that this achievement is impressive enough in its own right. The inventors of these strategies and machines have no cause to feel sheepish about what they have accomplished. The same thing applies, most emphatically, to the advances secured in recent years by those who study the real brain. Progress has been made across a broad front, contributions having been made by researchers in many disciplines.

We saw earlier that Turing's concern for the need to define intelligence led him to devise an imaginary experiment. The question of consciousness has naturally also given rise to much conjecture, one philosophical (and solipsistic) favourite being the difficulty of proving that one's fellow human beings actually possess it. For example, I know of no method whereby I could demonstrate to the reader of this book that it has not merely been written by an unconscious machine. Similarly, I would be at a loss to devise some testing procedure by which I could unequivocally establish the presence of consciousness in another person. The best that one can settle for appears to be an argument based on similarity. According to this prescription, which is in keeping with Ludwig Wittgenstein's philosophy, an individual who is aware both of his own consciousness and of his general resemblance to other people simply ascribes consciousness to them too. This is expedient, though it unfortunately proves nothing.

The difficulty of establishing the presence of consciousness in another individual is certainly not just a philosophical inconvenience. If that person is a patient on the operating table, and the surgeon wishes to verify *lack* of consciousness before getting down to work with his instruments, the issue acquires an obviously serious dimension. John Kulli and Christof Koch quote

from a number of case reports in which patients actually retained consciousness during surgery.[5] These unfortunate people were all pharmacologically paralysed while under anaesthesia, but they were conscious. Their various accounts of the experience make for chilling reading. As one of them put it:

The feeling of helplessness was terrifying. I tried to let the staff know I was conscious but I couldn't move even a finger nor eyelid. It was like being held in a vice and gradually I realized that I was in a situation from which there was no way out. I began to feel that breathing was impossible and I just resigned myself to dying.

Such failures of anaesthesia are not the fault of the surgery staff. Their guidelines are clear enough, and the routine tests of the effectiveness of what they administer usually produce an unequivocal result. The awful thing is that in some cases these chemicals simply do not have the desired effect.

Given the fact that identification of consciousness can sometimes be a very serious matter, it comes as quite a disappointment to find that even those who spend much time thinking about the question are still uncertain as to what consciousness actually is. A particularly disheartening note is struck in a dictionary entry by Stuart Sutherland.[6] It ends with the sentences:

Consciousness is a fascinating but elusive phenomenon; it is impossible to specify what it is, what it does or why it evolved. Nothing worth reading has been written about it.

Others do not throw in the towel quite as readily as this, but their attempts at a definition frequently betray an air of exasperation over the issue. Daniel Dennett, for example, seems overwhelmed by the plenitude of the phenomenon.[7] He has stated that:

Consciousness is both the most obvious and the most mysterious feature of our minds. On the one hand, what could be more certain or manifest to each of us than that he or she is a subject of experience, an enjoyer of perceptions and sensations, a sufferer of pain, an entertainer of ideas, and a conscious deliberator? On the other hand, what in the world can consciousness be? How can physical bodies in the physical world contain such a phenomenon?

These typical examples convey the nonplussed feeling experienced by so many writers on the subject.

There are two points that can be made immediately about Dennett's views. One is that it is not unusual for physical bodies to display phenomena that appear to be non-physical. The conduction, by solid objects, of both heat and electricity used to seem quite mysterious, and these properties were earlier taken to be the manifestations of ethereal processes. Now that such conductivities have been given physical explanations, however, they appear rather commonplace. And this example is a relevant one, because the transmission of signals to and from the brain, and also within the brain itself, is known to be mediated by a somewhat similar (electrochemical) type of conduction. The

other question that Dennett's opening remarks raise concerns the implicit assumption that consciousness is a product of the mind. One could ask whether the reverse is not closer to the truth.

This is not to say that Dennett is alone in adopting his position. The *Oxford English Dictionary* defines *mind* as being *the seat of consciousness, thought, volition and feeling*. The trouble is that the entry for *consciousness* merely sees this as being *the totality of a person's thoughts and feelings*. Between them, therefore, the two definitions seem to squeeze consciousness into a subsidiary position, the implied main connection being between mind, on the one hand, and thoughts and feelings on the other.

This sort of uncertainty, over the most basic aspects of the matter, suggests that our misunderstandings might be of a fundamental nature. The solution to the consciousness problem might be right under our noses. Our failure to realize this may be due to the fact that we have cluttered up the picture with too many unnecessary complications. Dennett's inventory of things mediated by consciousness underlines the richness of mental processes, but trying to cope simultaneously with all these, not to mention the many other manifestations of the brain's activities, might be counterproductive. Perhaps we should rather aim at paring the issue down to its bare essentials.

One could start by considering a new-born baby. It seems reasonable to accord consciousness to the wakeful neonate. But at its tender age, it must have a mind that is rather blank. During its first few moments, indeed, dangled upside-down by its ankles and bawling at the top of its lungs, its every response would seem to be the product of instinct rather than thought. My belief that this is the case draws strength from the lack of variation in infants' responses in this situation; one characteristic of the mind, on the other hand, is that it imbues its possessor with the ability to react in a non-standard manner.

I have chosen this initial illustration in order to emphasize the need to separate consciousness from its products. Even at this early point, I am thus advocating a somewhat different view from the one expressed by Dennett. I believe that the passage cited above *does* have things back to front, indeed, and that *it is the mind that is the product of consciousness*. I believe, moreover, that it is the sheer abundance of experience mediated by consciousness that fools us into misunderstanding the nature of this fundamental attribute.

In this book, I will probe the question of how consciousness arises from the brain's anatomy and physiology. A description of the workings of the brain's neural machinery will thus be an essential preliminary. Discussion of such details is unavoidable, if the arguments are to be more than mere hand waving. And an appreciation of these essentials will enable the reader to judge the merits of a number of suggestions. There will be conjecture that it might be possible to monitor consciousness externally, for example, and there will be speculation about the prospect of accounting for the biological determinants of intelligence, which enable an individual to reap the benefit of prior experience. I will also be discussing the physical mechanism underlying thought itself.