1

Artificial Brains or Artificial Experts?

My subject is *knowledge science.* It is the study of what communities know and the ways in which they know it. Individual human beings participate in knowledge communities but they are not the location of knowledge. Rather, the way that individuals reflect the knowledge of communities is a topic for analysis within knowledge science.[1] Knowledge science looks at how knowledge is made, maintained, disputed, transformed, and transferred.[2] Artificial intelligence is a natural field site for knowledge science because intelligent computers appear to channel and constrain what is known by knowledge communities into well-defined, discrete, asocial locations. Though early claims were overambitious, there are intelligent machines and they are getting better. Yet the existence of any intelligent machine seems to contradict a basic premise of knowledge science because a machine is not a community or a member of society. What better starting point could there be?

The early misplaced confidence of the proponents of artificial intelligence is easy to understand. It is precisely analogous to the misplaced confidence of rationalist philosophers of science and, I suspect, has fed upon it.[3] If science was, at heart, a logical, individualistic method of exploring the world, then the computer, a quintessentially logical individual, could start with arithmetic, graduate to science, and eventually encompass much of human activity. In the last two decades, however, science has started to look rather different. Detailed empirical studies of the way scientists make knowledge have given us a picture of science that is equally far from philosophical and common-sense models. Building scientific knowledge is a messy business; it is much more like the creation of artistic or political consensus than we once believed. The making of science is a skillful activity; science is an art, a craft, and above all, a social practice.

Starting from this viewpoint, the prospect seems distant of making intelligent, problem-solving machines. If science, the paradigm case of human problem solving, turns out to be messy, crafty, artful, and essentially social, then why should tidy logical and isolated machines be capable of mimicking the work of scientists? Still less should they be capable of doing the more obviously messy work of the rest of us.

The history and sociology of scientific knowledge has shown that scientific activity is social in a number of ways. First, when a radically new experimental skill is transferred from one scientist to another it is necessary that social intercourse take place. No amount of writing or talking on the telephone appears to substitute for visiting and socially rubbing up against the person from whom you want to learn. We can contrast two models of learning: an "algorithmic model," in which knowledge is clearly statable and transferable in something like the form of a recipe, and an "enculturational model," where the process has more to do with unconscious social contagion. The algorithmic model alone cannot account for the way that scientific or other skills are learned. My own study of the transfer of knowledge among laser scientists illustrates this point; I found that scientists who tried to build a radically new type of laser—a TEA-laser—while working only from published sources were uniformly unsuccessful (Collins 1974, 1985).

A second way in which science is social is that conclusions to scientific debates, which tell us what may be seen and what may not be seen when we next look at the world, are matters of social consensus. Whereas the formal model of seeing—the pattern recognition model as we might call it—involves recognizing what an object really is by detecting its distinguishing characteristics, the enculturational model of seeing stresses that the same appearance may be seen as many things. For example, there is a well-known photograph that can be seen as the face of Christ or Che Guevara, whereas it is claimed to be a picture of a snow-covered mountain range in China. Sometimes viewers see it as quite other things including an abstract black-and-white pattern of splotches. The question "What is it really?" cannot be answered. No amount of ingenious pattern recognition programming would reveal the truth. There is no algorithm for recognizing the pattern.

Nevertheless, in saying that it can be seen as an image of Christ or Che Guevara, I am saying something true about how our culture sees the image. For example, in the West it is easy to persuade people in a classroom that it really is a face. Students who cannot see the face come to believe that it is their fault. They feel inadequate for not being able to see what their colleagues can see so clearly. Using routine classroom techniques (Atkinson and Delamont 1977), a group of students can be made to act like a small scientific community. The nearest analogy is the historical and continuing debate about what is real and what is artifact when you look through a microscope. Like the face, scientific facts do not speak for themselves. Disputes in science are not settled by more and more careful observation of the facts; they are settled by broad agreement about what *ought* to be seen when one looks in a certain way at a certain time and location. Thereafter, anyone who looks and does not see what everyone agrees ought to be seen is blamed for defective vision (or defective experimental technique). This process is illustrated in earlier work on the controversy over the detection of gravitational radiation (Collins 1985 and a host of other detailed field studies).[4]

A third way in which science is social is in what one might call the routine servicing of beliefs. An isolated individual, having no source of reference against which to check the validity and propriety of perception may drift away from the habits of thinking and seeing that make up the scientific culture. Again, the social group is the living reminder of what it is to think and act properly, correcting or coercing the maverick back onto the right tracks. Thus, learning scientific knowledge, changing scientific knowledge, establishing scientific knowledge, and maintaining scientific knowledge are all irremediably shot through with the social. They simply *are* social activities.

What is true of scientific knowledge has long been known to be true for every other kind of human cultural activity and category of knowledge. That which we cannot articulate, we *know* through the way we act. Knowing things and doing things are not separable. I know how to speak through speaking with others, and I can show how to speak only through speaking to others. Changing the rules of speaking is a matter of *social* change; it is a matter of changing common practice. If the rules of speaking change, then I follow along with the others, not because people tell me what to do but because in living with others—in sharing their "form of life" (Wittgenstein 1953)—I change with them. I will change what I know about how to speak, not as a matter of choice, not as a matter of following a consciously appreciated rule, not at the level of consciousness at all, but because in doing what others do I will find that I know what they know. In knowing what they know, I will do what they do. This is true of speaking, writing, plumbing, plastering, practicing medicine, and discovering subatomic particles. To put the issue in its starkest form, the locus of knowledge appears to be not the individual but the social group; what we are as individuals is but a symptom of the groups in which the irreducible quantum of knowledge is located. Contrary to the usual reductionist model of the social sciences, it is the individual who is made of social groups.

Now think of a computer being tested for its human-like qualities. Let it be subjected to a "Turing Test" (see especially chapters 13 and 14) in which it must engage in written interchanges so as to mimic a human. The computer is at the wrong end of the reductionist telescope. It is made not out of social groups but little bits of information. What will it not be able to do by virtue of its isolated upbringing? Consider a more familiar example.

A foreign agent, of the kind one sees in the movies, has to pass a kind of Turing Test. Imagine a spy, a native of London, who is to pretend to be a native of, say, Semipalatinsk. The agent has learned the history and geography of Semipalatinsk from books, atlases, town guides, photographs, and long conversations with a defector who was himself once a native of the town. He has undergone long sessions of mock interrogation by this defector until he is word perfect in his responses to every question. His documents are in order and he has a story that explains his long absence from the town. In the films, the British agent goes to the USSR and begins to spy. He is picked up by the KGB and interrogated; the value of all those hours of training are revealed as he answers his captors' questions. As in the Turing Test, the problem for the interrogators is to distinguish between real accomplishments and an imitation—between the spy and a real native of Semipalatinsk. The moment of crisis occurs for our hero when an interrogator enters who is himself a native of the town. At this point nothing will save the spy except a distraction, usually extraneous, which brings the interrogation to an end. However good his training, we know that the spy will not survive cross-examination by a native of Semipalatinsk.

The reason we know he will not survive that final cross-examination is that, however long the spy's training, he cannot have learned as much about Semipalatinsk as a native would have learned by living there. There is a very great deal that can be said about

Semipalatinsk, and only some of it can have been said during the training sessions. The spy will be able to make some inferences beyond what he has been told directly (for example, he will be able to form some brand-new sentences in the language), but he will not have learned enough to make all the inferences that could be made by his native trainer or his native interrogator. A native learns about Semipalatinsk by being socialized into Semipalatinsk-ness and there is much more to this than can be explicitly described even in a lifetime.[5] Thus the trainer, competent native that he was, cannot have completely transferred his socialization to the spy merely by talking to him for a fixed period. Photographs and films will help, but all of these are merely different abstracted cross sections of the full Semipalatinsk experience. Willy nilly, the trainer must have talked about only a subset of the things he could potentially describe, and even if the spy has absorbed all his instructions perfectly, he cannot know everything that the trainer knows, nor everything that the native interrogator knows.

The native interrogator will ask questions based on his own socialization—again, he can only ask a small set of the potential questions—but there is a good chance that during the course of a long interrogation he will ask a question the answer to which covers details that the spy has not encountered, or turns on an ability to recognize patterns that he has not seen, or requires an inference that he is not in a position to make. Is there an area of the town—near the river, perhaps, or going toward the forest, or just beyond the tanning factory—that is quite distinctive to a native, but the distinctiveness of which cannot or has not quite been put into words or cannot quite be captured even in films and photographs? Is there a way of speaking or manner of expression or a way of pronunciation that we do not know how to document or that can only be "heard" as a result of very long experience with many native speakers? The interrogator might ask the spy to show him how the Semipalatinskians pronounce a certain word—not just tell him, but *teach* him—correcting minor errors as he does it. All teachers know just how hard it is to disguise book-learning for practical experience when confronted with an experienced pupil. These are the ways that the spy will be caught out.[6]

The TEA-laser study referred to above (Collins 1985) showed how laser scientists who had learned their craft solely from printed instructions were equally caught out. In that case they were unmasked when their TEA-lasers failed to work. The general rule is that we know more than we can say, and that we come to know more than we can say because we learn by being socialized, not by being instructed. The unspoken parts of knowledge are a different sort of commodity to the spoken parts: they are of a different substance, they have a different *grammar.* For example, just as these things cannot be deliberately told, even with the best will in the world,[7] neither can they be kept secret from a visitor to the society. It is not possible to imagine the whole population of Semipalatinsk starting to act like Londoners in order to prevent a stranger from picking up their ways of being.

If it is correct, this way of thinking about knowledge has significant implications for the future of intelligent computers; it will not be possible to construct the equivalent of a socialized being by giving a computer explicit instructions. On the other hand, if socially competent machines can be built without the benefit of socialization, social scientists will have to think again; if computers are unsocialized, isolated things, and if knowledge is as social as neo-Wittgensteinian philosophy would have it, then computers ought *not* to be able to become knowledgeable. Something is wrong. This argument applies as much to arithmetic as to spying. How can there be *Machines Who Think,* as the journalists put it (McCorduck 1979), unless they are also "Machines Who Live"—that is, machines who live with us and share our society?

Some optimists believe that machines who think are just around the corner, even though machines who live are still to be found only in science fiction. How can this be if the argument about the social embeddedness of knowledge is valid? How can the argument about the social embeddedness of knowledge be true, with all its implications about the cultural specificity of human behavior, if there can be intelligent machines? The artificial intelligence experiment is, then, not just a problem of engineering or psychology but an empirical test of deep theses in the philosophy of the social sciences. The possibility of a science of mankind, which emulates rationalist models of natural science, is being tested in attempts to build intelligent machines. If there can be machines that act indistinguishably from us, then the philosophical distinctions between action and behavior, and the argument about the peculiar nature of human rule-guided action, will turn out, after all, to be of no significance for the prospects of a *science* of society. The pigeons of philosophy of social science are coming home to roost in the intelligent computer (Collins 1985).

We must explain how it is, in spite of the social nature of knowledge, that we have machines that work. The explanation is not simple. First, machines that work rarely do the same work as humans. I will spend a lot of time showing that this is so. Second, machines that don't work can sometimes look like machines that do work. This is because these machines interact with humans who make good their deficiencies. Third, there is a large aspect of human behavior that mimics machines, and machines can mimic these aspects perfectly. To understand what is going on, we do not need to examine the ever-receding frontiers of computational science, we need only to think hard about machines that already live with us.

Machines That Work

Ironically, the journalists and computer scientists who say that machines that think are just around the corner have been too timid. The paradox is already with us and has been with us for a very long time. To the extent that we are prepared to be puzzled by the next generation of thinking machines we should already be puzzled by some existing computer programs, by pocket calculators, by slide rules, by logarithm tables, and by the printed word. In an important sense, all these are machines that work, and they work in a way just as puzzling as the machines that are more easy to anthropomorphize.

To start at the top, there exist computer programs that, given a spectrum of light, can work out the mixture of chemicals that produced it. Once upon a time, such chemical analyses were pieces of frontier science, and the proper interpretation of spectra must once have been the subject of dispute. Just as in today's contested sciences, this dispute must have been settled by forming social alliances, by forcing people to change their minds, and by breaking their hearts. For a human, learning to interpret chemical spectra still requires an apprenticeship—a period of socialization in chemical society. How can this analytic computer program work, then, without the equivalent socialization?

There are more famous and well-known examples of clever computers. By now almost everyone must have heard of Weizenbaum's ELIZA program (Weizenbaum 1976). Even the Apricot microcomputer that I am using to write this book can cope with ELIZA. For example, it engaged me in the following interchange:

HOW DO YOU DO, PLEASE TELL ME YOUR PROBLEM.
I'm feeling lonely.
IS IT BECAUSE YOU ARE FEELING LONELY THAT YOU CAME TO ME?
Yes, that's right.
YOU SEEM QUITE POSITIVE.
I just want someone to talk to.
WHY DO YOU NEED TO TELL ME YOU JUST WANT SOMEONE TO TALK TO?
Isn't it obvious?
I'M NOT SURE I UNDERSTAND YOU FULLY?
I need someone to care.
WHAT WOULD IT MEAN IF YOU HAD SOMEONE TO CARE?
I would feel happier.
CAN YOU ELABORATE ON THAT?
I need someone.
WHY DO YOU WANT SOMEONE?
etc.

ELIZA is meant to mimic a nondirective psychotherapist. It is a very simple program; nothing more, according to its inventor, than a programmer's parlor trick, yet it has been taken seriously. Weizenbaum relates how his secretary became engaged in a consultation with ELIZA of such personal significance that she refused to allow him to look at the computer screen. A number of psychotherapists, impressed with ELIZA's abilities, proposed that such machines had the potential to replace real doctors at greatly reduced cost. Yet ELIZA had no training in psychotherapy, nor had it the sort of wide socialization that one would expect to be a prerequisite of someone, or something, that could help others through their emotional crises.

Another program that achieved substantial success was Winograd's SHRDLU. SHRDLU appeared to be able to converse sensibly about a "micro world" containing blocks of different shapes and colors that it could manipulate. The world existed only within the program itself, but it was translated into an image on a monitor. SHRDLU could be instructed to move the blocks about, put them on top of each other, and so forth, and then answer questions in ordinary English about their dispositions. The following is a short extract of SHRDLU's performance taken from Winograd and Flores (1986, p. 109):

Pick up a big red block.
OK.
Grasp the pyramid.
I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Find a block which is taller than the one you are holding and put it into the

box.
OK.
. . .
What does the box contain?
THE BLUE PYRAMID AND THE BLUE BLOCK.
What is the pyramid supported by?
THE BOX.
etc.

Once more, this program appears to be holding the sort of conversation that would otherwise only be within the purview of someone who was thoroughly imbued with the culture of solid geometry and the English language.

But the mystery is even more intriguing than these examples suggest. Complicated, clever, state-of-the-art programs make us think that there is something very cunning in the program that enables it to interact with such apparent social accomplishment. The true depth and universality of the problem, however, appears only when we look at much more simple examples. Take a very simple computer—a pocket calculator; this seems to do arithmetic better than anyone I know. Yet doing arithmetic is again a skill learned through socialization and classroom practice. Even this quintessentially intellectual activity is learned within a community; we learn the *language* of mathematics. How can it be that my pocket calculator knows a language when it has never lived outside my pocket? Why stop at calculators? What about my slide rule? There is a sense in which it too can do arithmetic—certainly it and I can do arithmetic together—so it again must be partaking of the language of mathematics; is there a book to be written called *Slide Rules Who Think?* Yet my slide rule is not a social being. Are my logarithm tables social? They speak the language of mathematics with me in the same way as my slide rule.

We need not stop at the language of mathematics. The puzzle of computers, "How is an apparently social activity emulated by a socially isolated artifact?", is the same puzzle as how the printed word can carry knowledge between one person and another. All language is a social activity—how can it be encapsulated in inanimate paper and print? It is interesting that writing was once greeted with the same suspicion as expert systems are now. In Plato's *Phaedrus,* Socrates says:

> It shows great folly . . . to suppose that one can transmit or acquire clear and certain knowledge of an art through the medium of writing, or that

> written words can do more than remind the reader of what he already knows on any given subject. . . . The fact is, Phaedrus, that writing involves a similar disadvantage to painting. The productions of painting look like living beings, but if you ask questions they maintain a solemn silence. The same holds true of written words; you might suppose that they understand what they are saying, but if you ask them what they mean by anything they simply return the same answer over and over again. Besides, once a thing is committed to writing it circulates equally among those who understand the subject and those who have no business with it; a writing cannot distinguish between suitable and unsuitable readers. And if it is ill-treated or unfairly abused it always needs its parent to come to its rescue; it is quite incapable of defending or helping itself. (Hamilton 1973, l. 275)[8]

For Socrates, writing is but a pale shadow of social interaction.

The Social Nature of Artificial Intelligence

We have reached a point whence it is hard to see how to go on. Perhaps the social, enculturational model is wrong. Perhaps, while it is true that socialization is necessary for learning and transfer of knowledge, computers work because knowledge can be stored in a passive form within an isolated machine. There is another way of thinking about knowledge that makes it seem very much the property of the individual rather than the property of the social group. If I lock myself up in a room for a day, so that I have no contact with anyone else, when I come out in the evening my knowledge is not much changed. If it was true that I could speak English but not Chinese in the morning, in the evening I would still be able to speak English but not Chinese. Barring the possibility that I was in some form of extrasensory communication with English-speaking colleagues during the day, it looks as though all that social knowledge was fixed in my head the whole time I was in the room.[9] From this point of view, a facsimile of my head and body constructed during my day of isolation would have all my knowledge without ever being socialized or ever encountering another human being. It looks, then, as though one way to make a perfect intelligent machine would be to take an ordinary human and put him or her through a "Matterfax." This is a device, like a three-dimensional photocopying machine, that replicates the physical structure of matter down to the position of the last electron and the last quantum state. Given the Matterfax, there is nothing in principle to prevent knowledge being transferred to a computer.[10]

From this point of view, the problem of artificial intelligence seems to be about getting the same sort of complexity into a machine as is found in the brain. But this cannot be the problem we are dealing with here. I have argued that the conundrum is essentially the same for computers as for books; they too seem to mimic human linguistic capacity. Therefore it cannot be a matter of complexity. It is not as though a much more complex book will do the trick that a simple book cannot manage, and it is not as though a book mimics the content of the brain. The question we are dealing with is more modest. We want to know how things like books manage as well as they do in their interactions with us given that they are so far in substance and appearance from a Matterfaxed human being. And we want to know if extensions of our current methods of making books, and more intelligent artifacts, will lead us toward the Matterfaxed-style intelligent being by continuous incremental steps that we can foresee. We will not be able to understand how books and pocket calculators do so well by comparing what they do with the content of the brain. A book and a human brain are just too different for this to make any sense. To make progress in this direction we need to ask the question about artificial intelligence in a different way—a way that acknowledges the essentially social features of intelligence. The way that machines, or other simpler artifacts, fit into social interactions should be our starting point.

Fitting in is not always a matter of fitting perfectly. Consider the question: "Can we make an artificial heart?" By that question we mean: "Can we make a heart that will keep someone alive if it replaces his own heart?" The heart is not judged by reference to its own performance but by reference to the performance of the organism in which it is embedded. Suppose we made a heart that was slightly less efficient at pumping blood to the lungs than a real heart, but suppose the body responded to this marginally inefficient implant by producing more red blood cells so that the net amount of oxygen transported by the blood remained the same without any other disadvantages? We would consider this a highly satisfactory artificial heart even though the heart itself did not mimic the original. The same applies in a more minor way to the appearance and composition of artificial hearts. Appearance and composition affect the wavelengths that are reflected when light filters through the body's walls, and they affect the distribution of heat within the chest. But, within limits, the body is indifferent to these marginal changes. An artificial heart that had input and output characteristics marginally different from a real heart, but that was indistinguishable from a real heart in terms of the externally visible working of the human body, would be counted as a fine machine. Thus, from an engineering point of view, the performance of artificial hearts isolated from the context of the human body is not germane.

The same applies to the question of intelligent machines. There are two different questions relating to artificial intelligence. First, there is the psychological question, which is concerned with modeling the processes of the human mind with a computer. For psychologists, the purpose of mimicking human beings with a computer is to learn more about the processes of human cognition. To answer the psychological question we would want to mimic the workings of the brain.[11] What I will call the engineering question of artificial intelligence is quite different. The engineering question is: "Can we mimic the inputs and outputs sufficiently well to keep the organism going, irrespective of whether the mechanism corresponds to the original?"

The crucial difference between an artificial intelligence and an artificial heart is the organisms within which they function. For an artificial intelligence the organism is *not* the human body. When we ask whether we can make an intelligent machine, the big mistake is to think that this is the same question as: "Can we make an artificial brain?" But no one wants to remove a human brain and replace it with something whose artificial nature will not be obvious from the outside (as surgeons want to do with artificial hearts). There are philosophical and psychological debates about whether a person whose brain had been replaced with an artificial substitute with identical inputs and outputs would still be the same person. This sort of debate was once current in the case of artificial hearts, but it is not the sort of question that concerns us here. The organism into which the intelligent computer is supposed to fit is not a human being but a much larger organism: a social group.

The intelligent computer is meant to counterfeit the performance of a whole human being within a social group, not a human being's brain. An artificial intelligence is a "*social prosthesis.*" In the Turing Test the computer takes part in a little social interaction. Again, when we build an expert system it is meant to fit into a social organism where a human fitted before. An ideal expert system would replace an expert, possibly making him or her redundant. It would fit where a real expert once fitted without anyone noticing much difference in the way the corresponding *social group* functions.

Thus in artificial intelligence the question that is equivalent to "Can we make an artificial heart?" is "Can we make an artificial human?" And, just as an artificial heart does not necessarily have to have identical input or output characteristics (including appearance) to a real heart, neither does an artificial human. The embodying organism may be indifferent to variations, or it may compensate for inadequacies. As we will see, this explains the competence of programs such as ELIZA. ELIZA is hopeless as a brain but, in the right social circumstances, acceptable as a human.

The artificial heart analogy can do a little more work before we leave it. The body has an immune system that rejects foreign materials. Other things being equal, to be accepted by the body, an implant has to be designed to fool an alert immune system. There is another approach, however. Luckily for transplant patients the sensitivity of the immune system is not fixed; it can be reduced by drugs. According to the state of the immune system, the same prosthesis might be treated as an alien invasion or as a familiar part of the body. In the same way the social organism can be more or less sensitive to artifacts in its midst; one might say that it is a matter of the alertness of our social immune system. To use a term from debates in social anthropology, it a matter of the extent to which we are charitable to strangeness in other peoples.[12]

The admirable trend in the debates of the 1960s and 1970s was to see things from the other's point of view and thus increase our tolerance and reduce our tendency toward rejection or imperialism. One of the things I will try to do in the last part of the book, however, is to make our social immune system more sensitive to mechanical strangers, for, just as reducing the sensitivity of the physiological immune system carries with it enormous costs for the body, reducing our sensitivity to mechanical invasion has costs too. To learn to recognize artifacts for the strangers they are we need to understand their limitations. This draws us to ask the ultimate engineering question: "Given maximally vigilant humans, not only disinclined to compensate for machine deficiencies but actively seeking them out, what will give the machines' true identity away?" In other words: "What aspects of human ability can't machines mimic?" It is the sort of question that gives point to works of science fiction such as *Invasion of the Body Snatchers, The Stepford Wives,* or

Blade Runner, or Isaac Asimov's stories of robots. The possibility of human simulacra is well beyond the scope of this book, but the question is even more philosophically intriguing and informative when asked in more immediately relevant and limited circumstances such as the Turing Test. How can one learn to spot the deficiencies of a machine when communication is restricted to teletype terminals?

One final point of clarification: some people have a principled objection to anyone who says that such and such a technical development is impossible. The future, it is said, cannot be foreseen. In a vacuous sense this is correct. We can think only about *foreseeable* extensions to *current* ideas; it is not a matter of prophesying the future of mankind. I will suggest that certain things are not possible, but I mean by this only that such things cannot be envisaged by extending current thinking—not that such things will never come to pass in unforeseeable futures. Thus, if I say that emulating such and such a human ability is not possible, I will not have taken account of the Matterfax Corporation. I say such and such a thing is impossible in the same way as I might say that it is impossible that we will ever be able to buy a skin cream from the pharmacy that will allow us to take holidays in comfort on the surface of the sun. Perhaps such a thing will come to pass—but not by incremental progress.