The Open University

Philosophy and the Human Situation

Artificial Intelligence

Tim Crane

In 1965, Herbert Simon, one of the pioneers of the new science of Artificial Intelligence, predicted that ‘machines will be capable, within twenty years, of doing any work that a man can do’. Over thirty years later, there still seems no chance that this prediction will be fulfilled. My question is: is this a problem in principle for Artificial Intelligence, or is it just a matter of more time and more money?

I’m interested in the philosophical issues which lie behind this question. For whichever way we answer the question, it seems that philosophical issues – about the nature of mind and thought – are raised. To say that Artificial Intelligence could, with more time and more money, eventually produce a thinking machine is to be committed to the idea that (in some sense) thinking is a mechanical process. Alternatively, to say that Artificial Intelligence could never produce a thinking machine is to be committed to the idea that there is more to thinking than a mere mechanical activity.

Notice that the question here is whether Artificial Intelligence could produce a thinking machine, not whether it will. There may be all sorts of reasons – political, cultural, moral – why thinking machines will not be built. But these are irrelevant to the truth of Herbert Simon’s claim. For we do not think that the reason Simon’s claim is false is because the political, moral and cultural climate was not right. On the contrary: in the first half of the 1980s the US Department of Defence spent approximately 500 million dollars on attempting to develop Artificial Intelligence projects (including, for example, a truck that could plot its own course and steer for itself). This staggering sum only serves to underline the fact that the apparent failures of Artificial Intelligence are not due to intellectual Luddism. No. The question is one of the very possibility of Artificial Intelligence.

In this area of inquiry, then, some philosophy is inescapable. What I want to do here is to explain what I see to be the main philosophical issues in the area, and how they might be resolved.

To begin with, we must ask, why would anyone think that a machine could think? Why does this question arise, and why do people think it’s important? I think it’s useful to put this question in the context of the history of western science, and the history of the philosophical concept of the mind. For the main motivation behind the idea that a machine can think is the idea that human beings – and therefore the human mind – are just biological machines. So if the mind is a machine, its principles can, in principle, be understood and replicated, given adequate technology.

Now the idea that the mind is a machine or mechanism derives from thinking of nature itself as a kind of mechanism. So to understand this way of looking at the mind we need to understand – in very general terms – this way of looking at nature.

The modern Western view of the world traces back to the ‘Scientific Revolution’ of the 17th century, and the ideas of Galileo, Francis Bacon, Descartes and Newton. A useful way of understanding the difference between the medieval and Renaissance world pictures on the one hand and the modern world picture on the other is by contrasting the different ways in which they understand explanations. The medieval Aristotelian method of explanation – in terms of final ends and ‘natures’ – was replaced by a mechanical or mechanistic method of explanation – in terms of the regular, deterministic behaviour of matter in motion.

In what we could call the ‘organic’ world picture of the middle ages and the Renaissance, inorganic things were conceived along the lines of organic things. Everything had its natural place, fitting into the harmonious working of the ‘animal’ that is the world. But with the mechanical world picture, the situation was reversed: organic things were thought of along the lines of inorganic things. Everything, organic and inorganic, did what it did because it was caused by something else, in accordance with principles that could be precisely, mathematically formulated. René Descartes was famous for holding that non-human animals are pure machines, lacking any consciousness or mentality: he thought the behaviour of animals could be explained entirely mechanically. Though he was perfectly willing to regard animals as mere machines, Descartes did not do the same for the human mind: he placed the mind (or soul) outside the mechanical universe of matter. But many mechanistic philosophers in later centuries could not accept this particular view of Descartes’, and so they faced their biggest challenge in accounting for the place of the mind in nature. The one remaining mystery for the mechanical world picture was the explanation of the mind in mechanical terms.

So what would a mechanical explanation of the mind be like? At the very least, such an explanation of the mind must demonstrate how the mind is part of the world of causes and effects – part of what philosophers call the ‘causal order’ of the world. Another thing the explanation must do is to give details of the generalisations which describe these causal regularities in the mind. In other words, a mechanical explanation of the mind is committed to the existence of natural laws of psychology. Just as physics finds out about the laws which govern the non-mental world, so psychology finds out about the laws which govern the mind: there can be a natural science of the mind.

That, in any case, is how I see the background to the late twentieth century interest in the question of thinking machines. But what sort of machine is the mind supposed to be? The standard view within the mechanistic science of psychology is that the mind is a computer. Much contemporary psychology – at least in the branches which deal with reason and thinking (‘cognitive’ psychology) – assumes as a working hypothesis that mental states and processes are computational: that is, states of a computer.

So, given that not all machines are computers, we can distinguish between the very general idea that the mind is a machine and the idea that the mind is a computer, a special kind of machine. If the mind is a computer, then the aim of Artificial Intelligence is to find the principles on which the mind works – to find the mind’s program – and then to duplicate this program in an artificial (non-organic) machine. The mechanical world picture will then be complete.

The first thing we need to know is: what is a computer? We need not concern ourselves here with the details of how modern computers work. I shall instead assume the following rough definition of what a computer is: it is a device which processes representations according to rules. A computer processes representations – that is, items which represent or stand for things – according to rules which require absolutely black-and-white, yes-or-no answers to questions which are, so to speak, ‘posed’ by the representations inside the machine.

These rules are known as ‘algorithms’ or ‘effective procedures’. An algorithm is a procedure for finding the solution to a problem, which satisfies the following two conditions.

The first condition is that, at each stage of the procedure, there is a definite thing to do next. Moving from stage to stage does not require any special guesswork, insight or inspiration, and, second, the procedure can be specified in a finite number of steps.

So for example, the methods we use for calculating long division are algorithmic. The methods we use for making friends, on the other hand, are not. There is no algorithm for making friends.

That is what the theory of computation means by ‘rule’, more or less. What about representations? Take a simple example of a computer, an electronic adding machine. It’s natural to say that an adding machine performs the addition function by taking two or more numbers as input and giving you their sum as output. But strictly speaking, this is not what an adding machine does. For whatever numbers are, they aren’t the sorts of things that can be fed into machines, manipulated or transformed. (You can't destroy the number 3, for example, by destroying all the '3's written in the world.) What the adding machine really does is to take numerals – that is, representations of numbers – as input, and gives you numerals, representations of numbers, as output. So the notion of computation depends on the notion of representation. In the words of the American philosopher Jerry Fodor: 'No computation without representation!'.

Sometimes computers are called information processors. Sometimes they are called symbol manipulators. In my terminology, this is the same as saying that computers process representations. Representations carry information in the sense that they 'say' something, or are interpretable as 'saying' something. That is what computers process or manipulate. How they process or manipulate is by carrying out effective procedures or algorithms.

One important aspect of this abstract notion of a computer and computation is that it doesn't really matter what the computer is made of. What matters to its being a computer is what it does – that is, what computational tasks it performs, or what program it's running. The computers we use today perform these tasks using microscopic electronic circuits etched on tiny pieces of silicon. But although this technology is incredibly efficient, the tasks performed are in principle capable of being performed by arrays of switches, beads, matchsticks, tin cans and even perhaps by the neurochemistry of the brain.

Equipped with a basic understanding of what computers are, the question we now need to ask is: why would anyone think that being a computer – processing representations systematically according to rules – can constitute thinking?

The question is not: can the human mind be modelled on a computer? For even if the answer to this question is 'yes', this could not show that the mind is a computer. The British Treasury produces computer models of the economy – but no-one thinks that this shows that the economy is a computer. We need to distinguish between the idea that a system can be modelled by a computer, and the idea that a system actually performs computations. A system can be modelled on a computer when a theory of that system is computable. A system performs computations, however, when it processes representations by using an effective procedure or algorithm.

So to say that thinking can be modelled on a computer is to say that the theory of thinking is computable. This may be true, or it may not. But even if it were true, it obviously would not imply that thinkers are computers. Suppose astronomy were computable – that is, that the behaviour of everything in the universe could be modelled on a computer – this would not show that the universe is a computer.

On the other hand, we should not be too quick to dismiss the idea of thinking computers. The philosophical idea that the mind is a computer derives from ideas which have their natural home in the development of computers themselves: the attempt to systematise human thought. And in fact, whether our minds are computers depends on just this: whether the underlying mechanisms of thought are mechanical and systematic in the appropriate way: whether, in thinking, we employ or manipulate representations according to rules or algorithms. This is a question about the mechanisms underlying human thought – mechanisms in the brain. It's a scientific question whether there are such mechanisms and what their nature is.

That is what the question 'are our minds computers?' comes down to. However, the question, 'could any computer ever think?' still remains. In 1950, Alan Turing published an influential paper called 'Computing Machinery and Intelligence', which addressed this question. Turing asked the question, 'can a machine think?'. Finding this question too vague, he proposed replacing it with the question: 'under what circumstances would a machine be mistaken for a real thinking person?' Turing devised a test where a person is communicating at a distance with a machine and another person. Very roughly, this 'Turing Test' amounts to this: if the first person cannot tell the difference between the conversation with the other person and the conversation with the machine, then we can say that the machine is thinking.

There are two possible negative reactions to the Turing Test. One is to say that if you fill out the requirements of the test to make it a realistic test for intelligence, a machine could never meet it. The other is to say that even if a computer could pass the test, it could only be producing a simulation of thinking, not the real thing. These two claims actually define the main philosophical criticisms of the idea of a thinking machine.

I shall concentrate shortly on the first line of response. This says that the reason computers cannot think is because thinking requires abilities that computers by their very nature can never have. Computers have to obey rules but thinking can never be captured in a system of rules, no matter how complex. Thinking requires rather an active engagement with life, participation in a culture and ‘know-how’ of the sort that can never be formalised by rules.

The dominant idea here is that thinking requires commonsense knowledge, and commonsense knowledge cannot be represented as a system of rules and representations. The reason for this is that commonsense knowledge is, or depends on, a kind of know-how. Philosophers distinguish between knowing that something is the case and knowing how to do something. The first kind of knowledge is a matter of knowing facts (the sort of thing that can be written in books: for example, knowing that Sofia is the capital of Bulgaria), while the second is a matter of having skills or abilities (for example, being able to ride a bicycle). Many philosophers believe that an ability like knowing how to ride a bicycle is not something that can be entirely reduced to knowledge of certain rules or principle. What you need to have when you know how to ride a bicycle is not ‘book-learning’: you don’t need to employ a rule of the form ‘if the bicycle leans over to the left, then you lean with it’. You just get the hang of it, through a method of trial and error.

According to Hubert Dreyfus, a prominent American philosopher who (under the influence of Wittgenstein and Heidegger) has defended this approach, ‘getting the hang of it’ is what you do when you have general intelligence too. Knowing what a chair is is not just a matter of knowing the definition of the word ‘chair’. It also essentially involves knowing what to do with chairs, how to sit on them, get up from them, being able to tell which objects in the room are chairs, what sorts of things can be used as chairs if there are no chairs around – that is, the knowledge presupposes a ‘repertoire of bodily skills which may well be indefinitely large, since there seem to be an indefinitely large variety of chairs and of successful ways to sit in them’. The sort of knowledge that underlies our everyday way of living in the world either is – or rests on – practical know-how of this kind.

A computer is a device that processes representations according to rules. And representations and rules are obviously not skills. If a computer has knowledge, it must be ‘knowledge that so-and-so is the case’ rather than ‘knowledge how to do something’. So if Dreyfus is right and general intelligence requires commonsense, and commonsense is a kind of know-how, then computers cannot have commonsense and Artificial Intelligence cannot succeed in creating a computer which has general intelligence. The obvious ways for the defenders of Artificial Intelligence to respond are either to reject the idea that general intelligence requires commonsense; or to reject the idea that commonsense is know-how.

The second response is more popular. However, even if all commonsense knowledge could be stored as a bunch of rules and representations, this would only be the beginning of Artificial Intelligence’s problems. For it is not enough for the computer merely to have the information stored; it must be able to retrieve it and use it in a way that is intelligent. It’s not enough to have an encyclopaedia – one must be able to know how to look things up in it.

Crucial here is the idea of relevance. If the computer cannot know which facts are relevant to which other facts, it will not perform well in using the commonsense it has stored to solve problems. But whether one thing is relevant to another varies as conceptions of the world vary. The sex of a person is no longer thought to be relevant to whether they have a right to vote, but a hundred years ago it was.

Relevance goes hand-in-hand with a sense of what is out of place or what is exceptional or unusual. Here is what Dreyfus says about a program intended for understanding stories about restaurants:

‘The program has not understood a restaurant story the way people in our culture do, until it can understand such simple questions as: When the waiter came to the table did he wear clothes? Did he walk forward or backward? Did the customer eat his food with his mouth or his ear? If the program answers ‘I don’t know’, we feel that all its right answers were tricks or lucky guesses and that it has not understood anything of our everyday restaurant behaviour.’

Dreyfus argues that it is only because we have a way of living in the world which is based on skills and interaction with things (rather than on the representation of ‘knowledge that so-and-so is the case’) that we are able to know what sorts of things are out of place, and what is relevant to what.

This point, it seems to me, undermines the idea that a computer can think simply by being a computer. The question I asked at the beginning was: is the failure of Herbert Simon’s prediction the result of some problem of principle for Artificial Intelligence, or is it just a matter of more time and more money? The arguments of Dreyfus suggest very strongly that there is a problem of principle here: if a computer is going to have general intelligence – that is, be capable of reasoning about any kind of subject-matter – then it has to have commonsense knowledge. The issue now for Artificial Intelligence is whether commonsense knowledge could be represented in terms of rules and representations. So far, all attempts to do this have failed. It’s hard not to draw the conclusion that thinking cannot simply be manipulating symbols. Nothing can think simply by being a computer.

However, I’d like to end on a more optimistic note. For this last point does not mean that the idea of computation cannot apply in any way to the mind. For it could be true that nothing can think simply by being a computer, and also true that the way we think is partly by computing. So my conclusion is that even if there is no chance of a computer being able to think, it could be true that part of the story about the way we think is by computing. That would be an issue for another day, and another science – but not Artificial Intelligence.