The Computational Theory of the Mind

In the last 50 years, digital computers have developed from room-sized behemoths which took a week to test a 3000 digit prime to hand held devices which can triangulate a GPS position to within an accuracy of a couple of metres in a few hundredths of a second. In the same half century the assumption that the human brain is just a fabulously complex computer has become so entrenched in popular culture that it is almost heretical to consider the alternatives. And this is not without good reason. Nothing we have discovered about the detailed behaviour of the 100 billion neurons that we presume are responsible for whatever thinking goes on inside our brains shows any sign that they do not obey the classical laws of physics; in fact, each neuron appears to be just a straightforward logic circuit with inputs and outputs, just like a digital one. (Admittedly a typical neuron may have several thousand inputs and outputs and it uses a multi-level coding system not just a binary one but there seems to be no reason in principle why the action of any individual neuron could not be mimicked precisely with a small microprocessor chip.) The implication of this is that if you could replace every neuron in my brain with a suitable integrated circuit, the resulting machine would do exactly what my brain does. It would *be* a brain and, presumably it would *have* a mind.

Now this is not just philosophical speculation. This is a proper scientific theory with profound implications and testable consequences and it imposes some impressive limits on what computational brains can achieve. For this reason (and others) it has come in for a lot of serious criticism. Let us have a look at some of the implications of the theory and how these implications can be used to attack the theory.

The Computational Theory of Mind implies that the human brain is only capable of carrying out tasks which can be performed by a Turing Machine.

Roger Penrose has argued on the basis of Gödel's famous theorem that since human brains are capable of proving certain mathematical truths which it is impossible for a Turing Machine to prove, human brains cannot be Turing Machines. If this argument is upheld, it would prove beyond doubt that the human brain is at least in some respect non-computational but although Gödel's theorem is accepted by the mathematical community, its relevance to the issue of the human brain is widely disputed. In any case, it is not universally accepted that human brains can, in fact, prove certain mathematical theorems which Turing Machines cannot prove since you can always arbitrarily add the supposedly unprovable theorem to the list of axioms which the Turing Machine is given and – hey presto – the theorem is proved.

A Turing Machine cannot *create* anything; it cannot even, for example, print out a random number let alone write a novel or compose a symphony. It follows that a computational brain could not do these things either.

The problem with this is that it is impossible to define what we mean by a random number let alone a 'novel' or 'symphony'. In any case, it would be easy to equip a standard Turing Machine with a random number generator and in the course of time I have little doubt that it *will* become possible, if we should so desire, to program a computer to write simple novels and compose acceptable music by using a random number generator and a huge bank of literary and musical phrases.

A Turing Machine equipped with a random number generator might be able to write a hack novel or a halfway acceptable symphony but it could not invent a *completely new* plot or compose something in a *completely new* style.

Now we are getting somewhere. Human brains do seem to be remarkably inventive (but not uniquely so). However, it is by no means obvious that a Turing Machine with a random number

generator is logically incapable of composing something completely new. And in any case, even if you propose that the human brain is, in some way, non-computational, you still have the problem of explaining how it can create something that was not created before.

A Turing Machine can solve logical problems but it doesn't *understand* what it is doing.

This is the essence of Searle's objection to the computational theory of mind exemplified by his famous 'Chinese Room' thought experiment. The trouble with this argument is that we have now moved too far away from what is scientifically testable because the concept of 'understanding' is not sufficiently well defined. If we define the concept, for example, as 'the ability to respond appropriately to complex queries by drawing together many different aspects of knowledge and data from many different sources' then it is clear the, while the individual in the room does not understand Chinese, the *system* (which comprises the subject plus the dictionaries and data bases which the room contains) does.

A Turing Machine cannot *experience* anything. A Turing Machine equipped with a spectrophotometer could distinguish red from green but it cannot experience redness or greenness

Things which we can experience like redness and greenness are called qualia and their existence has been hotly (and in my opinion fruitlessly) debated over the centuries by philosophers. Yes, I experience qualia but it is quite pointless to attempt to prove that other creatures and machines do or do not because the whole point about experiences is that they are unique to the individual who experiences them. They are completely subjective and hence outside the remit of objective scientific enquiry. It is perfectly consistent therefore for the computational mind theorist to maintain that a television camera pointing at a red rag is experiencing redness and that the cameras experience of redness is no more different from mine than yours is.

A Turing Machine cannot feel *emotions*, it cannot be *conscious* and it cannot make decisions of its own *free will*.

Now we are so far away from objective scientific enquiry as to make the objection meaningless. Emotions like love and anger are experiences which don't even have recognizable stimuli so how are we to tell whether or not a Turing Machine is capable of feeling these things? Maybe consciousness is an automatic by-product of any sufficiently complex information processing. And as for free will, who is to say that we have it anyway? In fact all the computational theorist has to say in the face of these objections is that, while there is no objective proof that a Turing Machine is incapable of feeling emotions, being conscious and exhibiting free will there is ample proof of the contrary proposition and it is sitting on your own shoulders!

So where does this leave us?

So far all we have shown is that the computational theorist has sound counter arguments to all the objections which have been raised to the idea; but this does not prove that the idea is correct, only that it is logically possible. We must consider whether or not the non-computational alternative is better, more elegant, more economical or more testable. But first we must make an important distinction between conscious and unconscious minds.

Surely we can all agree that our minds are doing something qualitatively different when we are awake from when we are fast asleep. It is not that the brain is inactive when you are fast asleep, anaesthetised or in a coma; the brain is still receiving inputs from the senses and sending signals to various muscles to keep you alive. It can even, sometimes, coordinate so many actions that a sleepwalker can suddenly awake to find themselves walking down the street in their pyjamas. But I have never yet met anyone who has fallen in love, experienced redness, composed a symphony or even written a story when they are fast asleep. (Yes, you can probably do all these things when you are dreaming but that only goes to show that the state of the brain when it is dreaming is far closer

to its conscious state than when it is fast asleep – a prediction which is confirmed by EEG measurements of brain activity and measurements of metabolic rates which are typically 25-44% lower in deep sleep than in REM sleep.)

It is extremely tempting therefore to suggest that the human brain (and maybe those of some other animals too) can operate in two qualitatively different modes – a computational mode when it is asleep and a non-computational mode when it is awake (or dreaming). If this is true, then we can have the best of both worlds. The computational theorists can continue with their researches on vision systems or learning behaviour without having to bother about answering questions like 'What is it like to be a bat?' or 'Do crabs feel pain?'. At the same time, psychologists can continue to use concepts like belief and depression without being pestered about where the 'belief centre' or the 'depression neurons' might be located.

Now it would be premature even to suggest what a non-computational theory of the (conscious) mind would look like (although I do not believe that it is premature to start looking for structures in the brain that might suggest one). We do, however, have quite a lot of data concerning conscious minds with which any successful theory will have to contend. Much of this has been derived from careful and ingenious laboratory experiments (eg the discovery of 'blindsight' etc.) but we should not ignore completely that other source of information – flawed though it often is – namely, introspection.

Let us consider one concept that has been adduced to explain the way we think:

Mentalese

Last night I was in a drowsy state, half asleep, half awake, when I became conscious of the sound of rain pattering on the window. Suddenly I awoke with a jerk and felt a shot of adrenaline in my stomach. 'Christ!' I thought 'I left my trousers on the washing line yesterday because they were still a bit damp. They will be soaked!'

There are several interesting things to say about this little story. First, or rather, lastly, it was my *conscious* (and in my opinion, non-computational) brain which made this deduction. (If I had been fast asleep and the rain had fallen on me through an open window, my unconscious brain might have issued commands to my limbs to cover my head but it would not have come to the conclusion that my trousers were soaked.)

Second, my thoughts came into my head in *English*. Now many have deduced from this fact that if I had never learned to speak (in any language) then I would not have been able to make the logical deduction that my trousers were soaked. I don't ascribe to this view, partly because of the next observation.

Third, the shot of adrenalin occurred *before* I formulated the thought in English. The significance of this is that the logical deduction had actually been completed *before* the results were translated into English. Without prejudging in any way what such a language might consist of, we can usefully say that the brain uses an internal language, rather like machine code in a computer, in which to carry out its computations. We can call this language *mentalese* if we like. It would, however, be a mistake to think that mentalese necessarily looks like English or indeed machine code. The line following robot in the warehouse which strays off its painted line is not thinking 'my left photocell has stopped firing, I must speed up the right wheel' in any recognisable language – it is just wired up to do the appropriate actions at the appropriate time. Nevertheless, when the photocell stops firing the machine enters a state which in mentalese we might label '`I-am-off-line`'. Similarly, when my brain becomes aware that it is raining it enters a state which we can label '`I-am-aware-that-it-is-raining`'. These states are sometimes called 'symbols' and it is postulated that the human brain can hold millions of such symbols (i.e. memories) at the same time.

All the computational theorist has to do now is to invent a kind of quasi-logic in which these

'symbols' in 'mentalese' are processed and the desired conclusion reached.

`'I-am-aware-that-it-is-raining' + 'My-trousers-are-outside' + 'Objects-left-outside-when-it-is-raining-get-wet' = 'My-trousers-are-wet'`

I am afraid this just won't work.

It is not that the logic is faulty – it is just that there just far too many alternative symbols that may have to be examined before the brain hits on the right one. Do we really think that the brain actually considers all the other possible symbols which it holds in its memory? Does it also consider the symbol '`My-trousers-are-grey`' and '`My-bicycle-is-outside`'? Why not also consider '`My-bicycle-is-red`'? And what about using the principle that '`Objects-left-outside-often-go-rusty`' instead of `'Objects-left-outside-when-it-is-raining-get-wet'` etc.etc. I do not believe that the brain does a Google-type search of all the possibilities. True – Google search algorithms are impressive, but the speed of the search engine is bought at the price of huge computing power and vast petabytes of storage holding indexes to improbable combinations of words just in case someone might ask one day.

But what is the alternative? Are we hard-wired like the line-following robot? Perhaps, when I hang the trousers on the line, certain synapses are strengthened and other connection made between the neurons that hold the `'Objects-left-outside-when-it-is-raining-get-wet'` symbol and the `'My-trousers-are-outside'` symbol so that as soon as the `'I-am-aware-that-it-is-raining'` symbol is activated the `'My-trousers-are-wet'` symbol lights up.

But in truth, this idea is little better because it is difficult to see how the neurons 'know' which synapses to strengthen and which connections to make. In order to strengthen the link between `'Objects-left-outside-when-it-is-raining-get-wet'` and `'My-trousers-are-outside'` the neurons must also 'know' that '`wet-trousers-cannot-be-worn`' and '`rain-causes-trousers-to-become-wet`' etc. etc. etc.

My conclusion is that the computational theory of mind is missing something essential – something we fundamentally do not understand. If our brains were entirely computational we would only be able to carry out tasks that we were pre-programmed by our evolutionary development to do and our capacity to learn new tricks would be limited to simple stimulus-response type behaviour reinforced by constant repetition. This is the behaviour exhibited in abundance by many species of animals but the behaviour of primates and many other mammals and birds is much richer than this.

Free will

I have another objection to the computational theory of mind which is concerned with the issue of free will. This objection is easily countered by simply denying the existence of free will and adopting a rigorously deterministic stance but I am extremely reluctant to do this. I simply have to believe that I have a choice in the matter of what I do next otherwise there is absolutely no point in discussing moral issues, interviewing for a job, putting money into a bank, shopping for tomorrows dinner or even taking another breath. I accept that a spider can build a web, a fish can find its spawning grounds, a bird can navigate across the ocean and a bull can mate with a cow by instinct alone without any need for the animal to make any (conscious) choices but I will not accept that I myself did (the human equivalent of!) all these things either unconsciously or as if I was controlled by external forces. It seems to me to be an inescapable fact that a being which has free will *must* be conscious and I think it highly likely that all conscious beings possess free will.

If this is true, then any explanation of the conscious mind which fails to predict or explain free will is incomplete.

But there are at least two serious objections to this. On the one hand you might claim that this

only makes the problem worse because it is even more difficult to explain free will than it is to explain consciousness. I disagree. Because consciousness is a subjective phenomenon, it cannot be studied scientifically (though its manifestation in the physical brain can). Free will, however, is characterised by certain kinds of behaviour which is easily studied. The second objection is that the classical laws of physics specifically rule out any behaviour in any system which is not completely deterministic and even quantum systems are either deterministic or random. There is simply no place in Physics for free will. So much the worse for the laws of Physics, I say.

A way out of the impasse

There is, I believe, a way out of this impasse. It is not that the laws of Physics as we currently understand them (and I am talking about quantum theory here) are wrong – it is just that we have not yet realised all their implications. We know for a fact that objects on an atomic scale do not behave in the way that large objects behave. Electrons can be in many places at once; two photons widely separated can nevertheless share properties; atoms can jump from one state to another without passing through any intermediate stages etc. etc. So far this behaviour is confined either to atomic objects or macroscopic objects at very low temperatures and the current accepted wisdom is that nothing in a human brain, let alone the whole brain itself, could possibly exhibit any kind of quantum behaviour. But the history of science is littered with examples of paradigm shifts in which accepted wisdom is discarded in favour of something new. Let us think the unthinkable and suppose that inside a conscious human brain some process is going on of which we currently have no knowledge at all. It is possible that this process is something like the entanglement which links two separated photons but another requirement is that it must extend over large parts of the conscious brain. So that we can talk about it, we shall give this process a name. Lets call it *macroentanglement*.

All that we require of this process is a) that the phenomenon of consciousness is a necessary product (or even by-product) of this process and b) that the process permits the conscious object (i.e. the brain) to influence the outcome of certain critical events (such as the firing of a neuron) which is not already uniquely determined by the laws of physics (e.g. a system in a quantum state which contains some undecided factors). We can now justifiably say that 1) the outcome was not pre-determined (because the quantum state permitted both the possible outcomes) *and* 2) the outcome was not random because the conscious object caused the one outcome and not the other. In short we can say that the conscious object exercised its *free will*.

Pinker's 5 problems

In 'How the mind works' Steven Pinker lists 5 baffling problems to which he admits to having no answer. I believe my theory answers three of them

Problem No. 1 – what is consciousness?

Pinker asks '*How could an event of neural information processing cause the feel of a toothache ...*' My answer is that the neural information process is, in fact, macroentanglement but macroentanglement does not *cause* consciousness – it *is* consciousness.

Pinker asks '*How could I know whether a worm, a robot, a brain slice in a dish, or you are sentient (conscious)?*' Answer: check to see if macroentanglement is going on.

Pinker asks '*Is your sensation of red the same as mine or might it be like my sensation of green?*' Answer: insofar as my macroentanglement is the same as yours, my sensation is the same as yours – but insofar as my sensation is mine and not yours, they are completely different. Stupid question.

Pinker asks '*What is it like to be dead?*' Answer: you really want to know? Here, let me shoot you. Except that you won't. Because you will be dead. (Who is this idiot?)

Problem No. 2 – what is the self?

Pinker asks: '*Say I let someone scan a blueprint of my brain into a computer, destroy my body, etc. etc.*' I don't have to know the rest of the question. As soon as the brain is destroyed, the self is destroyed as well. In fact the self is destroyed even when the brain goes to sleep. A self only exists (if it exists at all) when the process of macroentanglment is evident.

Pinker asks: '*When does a zygote acquire a self?*' Answer: only when its brain is sufficiently developed to employ macroentanglement

Pinker asks: '*How much of my brain tissue has to die before I die?*' At last, a reasonably sensible question. My answer is that it will all hinge on how much macroentanglement is going on. The issue is of crucial importance in the case of patient in a coma. If we ever get to the stage of being able to measure macroentanglement with some kind of quantum scanner, we will know for sure but at the moment, we just do not know whether certain patients are or are not consciously aware of their surroundings. Our best guesses are currently based on EEG traces and other types of brain scan and the whole science on anaesthetics is based on the premise that when the EEG traces show that the patient is asleep, we can assume that they have lost consciousness and that they are (temporarily one hopes!) brain dead.

Problem No.3 – do we have free will?

Pinker asks: '*How can my action be a choice for which I am responsible if they are completely caused by my genes, my upbringing and my brain state?*' Answer: there is some justification for maintaining that a homicidal maniac is not responsible for his actions because of his genes and/or his upbringing but there is no justification for for claiming that his brain state relieves him of responsibility since *he* **is** his brain state.

Problem No. 4 – what is meaning

Neither my theory of mind nor Pinker's sheds any light on the problems of epistemology.

Problem No. 5 – what is morality?

And the same is true of ethics.

Surely a theory which can illuminate three out of five major philosophical problems which have puzzled philosophers for at least 2 millennia has to be worthy of serious consideration?

J Oliver Linton

Carr Bank, December 2015