GRESHAM COLLEGE

21ST JANUARY 2020

CAN MACHINES BE CONSCIOUS, AND WOULD IT MATTER IF THEY WERE?

PROFESSOR YORICK WILKS

People continue to ask questions such as "Could a machine think?" and "Could a machine ever be conscious?" The last question used to be banned in philosophy, or at least in philosophy done in English, but in recent years it has come back with great force, and provoked much new discussion in the light of discoveries made in brain science and advances in Artificial Intelligence (AI) AI workers are, by and large, naive materialists and mechanists, which means that, for them, mental processes are no more than software running on hardware made of flesh. And saying that does not really need to be justified for them; it is simply an assumption that allows them to get on with their job of constructing mechanical analogues or simulations of ourselves, who are, in Minsky's memorable phrase, "meat machines".

Strong assumptions about an underlying mechanism like this are normal in the sciences: they are what allows experimental, as opposed to philosophical, work to proceed, and they are justified when the experiments turn out well. But AI is not really an experimental science at all, but an engineering technique or, if you want something more dignified, a practical task, a little like the alchemical tradition preceding chemistry, and its suggestiveness for investigating the nature of consciousness rests entirely on that fact.

The "normal" situation in the sciences is not a sure guide when the brain and mind are the subjects of investigation, which is largely because of the peculiar features that attach to the notion of consciousness, and to its close relations thought, feeling, privacy and so on. In this lecture, I shall play fast and loose with these words and the subtle distinctions between them. That is to say, I shall not distinguish, as carefully as some philosophers would, such questions as:

- Does a Computer think?
- Is a Computer conscious?
- Does a Computer have essentially private inner processes?
- Could a Computer be aware of a sensation of warmth?
- Is my Computer aware of those cushions in front of it?
- Is my Computer conscious it is telling me the temperature?

And so, will not be treating these traditional questions with the respect they have earned over the years. Nor will I provide you with pages of solid stuff about how we normally want to distinguish between a dog being conscious and unconscious, in the sense of asleep, and that the answer does not commit us to saying that the dog has consciousness, etc. You can find plenty of that done better by others elsewhere, and what I have to say can be put much more simply, and, at the end of the lecture, in terms of how computer models work.

When I write of consciousness, I shall mean what is usually called self-consciousness, rather than simply having sensations or perceptions. And I shall not try to define what kind of thing consciousness is, but just assume for the

moment that if you are reading this, then you know already, and we can build on that so as to ask what it could be like for something very different from you and me to have it. What we shall be trying to do in this lecture is to see whether the analogy between people and machines can give us a useful way of talking about consciousness, whatever it is.

It is important to see from the beginning how utterly opposed different conceptions of consciousness are: in the early 20C William James would write

"'consciousness' is the name of a non-entity and has no right to a place among first principles".

Although Freud was writing at about the same time:

"what is meant by 'consciousness' we need not discuss. It is beyond all doubt".

In recent times, the philosopher David Chalmers has claimed that consciousness is the hard problem, while Galen Strawson has argued that consciousness is the one thing we are most acquainted with and know most surely. These positions are a complete reversal from fifty years ago when such subjects could be discussed on the Continent in foreign languages but philosophy in English was concerned with the analysis of language and abhorred, with a few exceptions, everything that was totally private and could not be publicly examined and discussed: such as consciousness along with dreams, pain, and a language used by only one speaker.

The Continental Tradition About Consciousness Was Different

The discussion of consciousness in Continental philosophy has been quite different from that in the English speaking world, and was central to the philosophy of Husserl, for whom the notion of *attention* was key---for him consciousness is always *of* something, not everything, and is thus selective. In the 19C Hegel called consciousness *Bewusstsein,* which he saw as more than individual and something closer to a group or world consciousness, and which was the only conscious part of the whole universe. This tradition can be related further back to the individual consciousness in all things, what he called *monads*, in Leibniz's philosophy, to which we will return to later since monads are the origin of the notion of a closed "information module" in AI, and Leibniz is the main bridge from traditional philosophical thinking on these issues to AI, since he was also the first person to set out clearly the idea of a reasoning machine. The notion that there can be a group consciousness, beyond the individual person, is also central to the influential psychology of Jung and to religious thinkers like Teilhard de Chardin for whom a group human consciousness was something that was being evolved, even in our own times.

The idea that perhaps everything is conscious is usually called *panpsychism*, and there is a strand in modern philosophy, of which Galen Strawson is the best known representative, that claims that this view is a form of *idealism* (usually taken to mean that only the mental world is real) but is also *physicalism* as well: that the physical world is real and consists of consciousness, the most real thing there is. Those who hold this view often argue from the state of quantum mechanics in which certain phenomena in fundamental particles seems essentially mental and not independent of our consciousness in that way that standard physicalism would assume, when it takes the physical world to be quite independent of our existence.

There is relevance of all this to the state of AI and particularly to the world wide web: some would argue that the existence of the web of all human knowledge, that we now have in rudimentary form, is itself a form of that objective world consciousness of all human thought already coming into being, and in a way that can be related Hegel's Bewusstsein. Again, it could be argued that the idea---to be found in Aristotle and Leibniz---that all things are conscious to some degree would take that as implying that machines also must therefore be conscious to some degree, and them showing more of the behaviour of being conscious, in the way we are, is not in itself surprising. Those familiar with Japanese thought claim that these panpsychic ideas are to be found in traditional thinking there which is why Japanese culture is more accepting of robots in society that we are in the west.

Features of AI Programs

In this lecture I shall also present ideas drawn from AI programming, or more generally from Computer Science, where a mechanical analogue of consciousness, might be sought in the future. These will not be the obvious places and have not been subjected to much philosophical investigation. Here are four technical notions in computing which may help us with consciousness, improbable as that may seem at first:

- Modularity
- Implementation independence
- Program level reduction
- Program inference

These are not really technical terms, just convenient labels, though each is close to a well understood technical notion. They are four the places to start our search for features of modern machines that might suggest models of consciousness.

Modularity

Modern computer programs, especially those in areas like AI, are written as interconnecting sub-parts or modules, and not as seamless wholes. Modules do not have access to the contents of other modules but simply exchange information without, in Carl Hewitt of MIT's immortal words, "being able to dicker around with the insides of their neighbours". Herb Simon argued in l969 that evolution would almost prefer structures (in brains, presumably) that are decomposable into modules in this way, and that modularity may therefore be expected in "genetic programs". Simon's story to illustrate this point compares the commercial viability of two watchmakers, one of whom puts watches together out of finished sub-assemblies which cannot fall apart, and another who assembles each watch from all its basic parts and risks the whole thing falling to pieces if dropped. It is obvious that the first watchmaker, with watches assembled from "modules", will do better.

It was this general idea of modularity that Minsky had in mind at MIT when he has suggested at various times that an organism would be more efficient, in terms of its ability to survive, if it also had, as a separate module, *a "model" of itself*, one that could of course be totally false as to the facts of the self's own reality. But alcoholics who believe themselves to be alcoholics probably do survive better than those who believe themselves to be merely social drinkers. So, an accessible model of the self is clearly one of the first places that one might look for analogues of consciousness in a machine system.

Later, Minsky revived these notions, explicitly drawing analogies with the sorts of "modularity" to be found in the psychoanalytic writings of Freud, and in particular his famous location of entities called the Ego, Id and Superego in the human (conscious and unconscious) psyche. This assumption that Freud's is just another modular view of the psyche is probably unfair to the totality of his work, although it is certainly the popular view of him.

As I noted earlier, he original idea of a module goes back to the philosopher Leibniz's *monads* in the 17C, entities of which he thought everything was composed and which were, in varying degrees, conscious of the content of other monads and of themselves. A monad that was a human soul or psyche would have access to a nearby monad that was its "supreme organizer", which plays a role in the human psyche something like that played by God, the supreme monad, in the whole universe, who had access to the contents of all other monads. Minsky suggests that Leibniz' notion is not too far from that of a module with access to a model, or representation, of how it itself is related to all the other, lower, modules that comprise it, as well as comprising the whole world around it, as it were. It might also be expected to have a property close to what we ordinarily call consciousness or self-consciousness. This additional property, Minsky believed, could also have a functional or evolutionary explanation, of the sort suggested for the property of modularity by Simon, so that a potentially "conscious" supreme organizing module would therefore be in a position to "debug" or adapt relations between other modules.

These are very general ideas indeed, and it may still be difficult to make the connection Minsky wants us to, between common sense about traditional properties of consciousness (vague as they may be), and the notions of getting access to the contents of "other" modules or not being able to, fundamental as those are to any account of intelligent mechanisms. There are certain yogis in Asia who appear, on all the evidence we have, to be able to take control of some of their physiological functions (their heartbeat rate, digestion etc) which are completely inaccessible to most of us. They can, if these claims are true, debug, or at least change, their "digestive program", or even slow down their hearts, so the limits of our consciousness may not be absolute but adaptable with the right training.

Implementation independence

It is well-known that the same computer program can be run on different of machines, not only different individual machines but machines different types, and that extends even to machines working with quite different physical processes. This is what is meant by the implementation of a program being machine-independent, and it is part of the conventional distinction between hardware (i.e. machines) and software (i.e. programs), namely that software is, or can be, more or less independent of the hardware it runs on.

As always, things are not so simple, and all PC users know that Windows/PC machines did not initially run Mac/Apple applications, nor vice versa, because such programs depend on their particular hardware. But it is now generally agreed that hardware independence is a Good Thing, which accounts for the rise of "platform independent" programming language like Java, and Macs are now based on the, very general, Unix software system.

It is this portability aspect of programs, and the conventional hardware-software distinction that goes with it, that has most interested those concerned to explain the relation of brains to minds in computer terms. There has been an easy temptation to exploit the hardware-software distinction as a model of the brain-mind distinction. That leads directly to a portable notion of mind (or soul or spirit), quite independent of a particular body, one that many have always found attractive on theological grounds. Support for it has come from the observation that both the brain and the conventional digital computer (i.e. one hard-wired only for its machine code or very lowest "level" of language, a key notion we shall come back to) seem to be surprisingly homogeneous in their internal structure, which led to remarks like Newell's "..intelligent behaviour demands only a few very general features in the underlying mechanism".

We shall not make use here of this mind-brain analogy, because, as we saw, there is a conventional element in it: the hardware-software boundary can shift at different times and with different styles of machines and languages, since there have been machines that essentially copy program concepts directly into their hardware construction. Nonetheless, implementation independence has had a powerful effect on AI thinking about metaphysical problems, and was behind McCarthy's insistence over decades that AI must be defined as the study of intelligent mechanisms "independent of their implementation in machines or brains" which distinguishes it firmly from psychology which is always about people, and has the odd consequence that AI is not strictly *about* machines at all.

Program Level Reduction

This phrase captures a notion close to the last but concerns not the translation of procedures from programs/software to machines/hardware but the translation of procedures from one "level" of a programming language to another. In a digital computer the "lowest level" of language is simply a string of binary digits (1 and 0) that is identical to the states of the machine's hardware registers or switches. This binary language is a level below what is usually called "machine code": normally instructions to add, subtract or shift the contents of whole registers (themselves strings of binary numbers). At a level above machine code is "assembly language" whose commands normally translate into a set, of tens or perhaps a hundred, machine language commands. This ascent up the levels of programming languages can go on without any natural limit: at the very highest levels of language, are the old AI languages Lisp and Prolog. At every level, the language strings produce the same effects as the strings below, which is what we mean by them being translations of each other.

As one goes higher up this ascent of levels of languages, the code becomes progressively more like a natural language, such as English, though nothing particularly mysterious hangs on the fact and it does not mean that one day we will be able to write programs in English. Ordinary human languages are just not the best way of saying very precise things, but already one can say things like "raise the temperature five degrees" to a house heating system, and that will become increasingly common. In that case the heating system does indeed translate that English sentence down through levels of code to a machine code that changes the thermostat, so one could say, in that example, that the English command was just a very high-level piece of program. When automated driverless cars appear in the near future, we will be able to jump into one and just say "Take me to Wigan" in English and it will.

Whenever a program is executed, a sequence of such translations between levels of language is carried out, but the upper levels of the program have no direct access to the levels below them, and the programmer who writes at the topmost, or accessible, level has no need at all to know how his program is translated when the program runs.

This phenomenon is very suggestive of a feature of conscious experience, especially our lack of conscious access to how we do what we do with our bodies. When someone says, "We went to a bar and ordered a drink", everyone agrees that the sense of "bar" here is a drinking place and not a rod of iron. Linguists, psychologists and AI researchers have theories about what procedures might select the right sense and reject the wrong ones. But the speaker of that sentence has no idea at all how he does it and may well have a healthy scepticism about whatever a theorist tells him about what he is "really" doing. Yet how are we to understand this situation: speaking like that is the product of rules (for it is certainly not done randomly) and yet the performer has no access whatever to these rules? We could say something similar about how one breathes without knowing ---and has done for all human history----how knowing the biochemistry of how one's lungs use oxygen.

Program Inference

This phrase covers limitations on our ability to see in the "other direction" from the one we have been discussing: that we cannot infer the highest level of a program in a system given the lowest. If one stood in front of a large old-fashioned computer from the Sixties—the kind that is still used in cartoons----one would see banks of lights flashing. These rows of lights are in direct relationship to key registers inside the machine and actually express the binary numbers in them (a light being off for 0 and on for 1) at that instant. If all the internal processes of the machine were slowed down, one could actually see a representation in lights, on the console, of each command being executed by the machine at the lowest possible level of language, that of binary numbers 1 and 0. Now, if on could read all those binary numbers in sequence, could one work out the highest level of the program or, to put it another way, could one infer what the machine was actually up to, in the sense of paying tax refunds to the citizens of York, as opposed to translating a book from English to Chinese? And, if there are limitations on our ability to work that out, are they serious or would they just require more effort than anyone is normally prepared to put in?

Such an ability ----if it exists----would be of great practical interest to many groups of people: there are specialists, detectives almost, who can take enormous quantities of program in a lower-level language (not binary numbers, but normally machine code or something a little "higher") and make plausible guesses as to what the program actually does at a higher level of description. These are the kinds of people who can take the chunks of Microsoft Windows source code, that was released by mistake onto the Internet recently, and work out what each bit probably does and how to use that information to write new and more scary viruses.

This "inscrutability" of computer code has certain obvious relations to the nature of the brain, where philosophers and physiologists have agreed over a long period that no mental contents-----what someone is thinking about—are to be discovered from examining the structure and firing of individual brain neurons. At the moment we cannot even imagine what a "brain programming language' would be like.

One thing we might expect of any machine that was to be considered conscious that it would have to have the sort of final authority over what state it was in that we normally concede to humans. When Mr. Jones, lying fully conscious on the neurosurgeon's table, insists that he is in pain, we have to believe him, even though the neurosurgeon says that, given the position of his brain probe at that moment, Jones should not be feeling a thing. We tend, naturally, to let people have the last word on what they are feeling.

If a computer told us that its memory was suffering from a certain kind of fault, we might be persuaded from past experience to go on examining its hardware for faults, even though we found none in the initially plausible places. We might, to speak anthropomorphically, allow it to insist that its memory was going and, if we did, we would be allowing it just that authority we normally allow only to people. We might come to agree that a computer really was paying tax refunds (because it said it was), even though all detective work on its machine code seemed to suggest it was occupied directing the trajectories of intercontinental ballistic missiles. If we did come to allow such authority, or privacy, to the machine itself, then huge consequences would follow, for its blueprints and machine programs would no longer be a safe guide to its future behaviour.

To return to the main question: does all this suggest anything insightful about the nature of consciousness? Given that preserving the privacy of consciousness is one of the conditions for being an analogy here between humans and machines, any interesting sense of machine privacy we can find support for should be relevant to this question. But someone might say privacy is unconnected with consciousness because, for example, when we are driving, the highest-level commands from the dive to turn the car etc., would also be "inscrutable from below", that is, from observing our brains while driving. But the top level of those commands could never be identified with consciousness---as we suggested the top levels of programming languages actually expressed "what the program was really doing" -----because, as Sartre used to point out when discussing phenomena like this, we can drive "without thinking about it" and at no danger to the public.

This is true, and we are not trying to identify top levels of programs with consciousness in any simple-minded way. The fact that consciousness can only be of the topmost level of brain activity, by definition, does not mean we are always conscious of that level. Clearly, in "unconscious driving" we are not.

Everything said here seems to depend on the assumption that there is only one thing a computer really is doing, just as for people, there is *one thing* they are doing, or which is in the immediate content of their thoughts. Both of these might seem just common sense, and that a computer is never both calculating tax and aiming missiles, just as people are always doing or thinking one main thing even if they manage to drive at the same time! About people, it may seem common sense that we are "only thinking about one thing at once", but about computers that seems just not true. What a computer is really doing can never be more than a matter of interpretation by some human users of a computer, and not a choice made by the computer itself, because it cannot do that. If a detective approaches someone and says "all your activities of the last week are consistent with you planning to rob a post office" then, when the person replies, "but as a matter of fact, I just am shopping and nothing else, and that's that", then one can either refer to intentionality , as writers like Searle would do, to mean that only the person really knows what he is doing and no one can question that. Or, like the detective, one could continue to keep the person under close scrutiny. But we may assume that the person them self is the final authority on what they are *really* doing.

But in the computer case, they do nowadays do many things at once: a PC can download mail while uploading one's photo collection on the same screen. It may be true that, at any given instant in the machine's registers, it is running code for only one of these things, but at the top-levels we are talking about and observing the screen, there is every appearance of the machine doing several different things at once.

It may be that computers are essentially different from humans in that respect, even though for both "real content and activity" at the highest level of coding are not discoverable from observing the lowest. Leibniz famously argued that, in terms of his monads that composed the world, God was the supreme monad that was aware of the contents

of all other monads at once. His was the only monad that was aware of many things at once or, in another metaphor Leibniz used, that could see things from every point of view at once. Some theologians have argued from this that God could not, on such a view, be conscious in the way we are, because consciousness, as we have argued throughout this lecture, requires some limitation, some *deprivation* of access. So, if computers really can do more than one thing at once "at their topmost level" and, if the language level is any kind of analogy of consciousness, then it must follow that computers, in so far as they could be conscious, would be a stage further on than humans towards the condition of Leibniz's God monad.

Must Consciousness be Evolved in Time?

One crucial aspect of the problem is that although most us have an intuitive feel for our own consciousness – though we cannot describe it well, except to say when we are losing it as we fall asleep or have a presurgical injection – we have no such direct knowledge that *anyone else* is conscious. As some philosophers like to remind us, other people could all just be zombies who claim to be conscious, just as chatbots do when asked, after they have been programmed to say that. Saying one is conscious proves nothing. In the lecture we shall look at recent proposals for tests that might be applied to see if a computer had, in any sense, become conscious.

"Becoming conscious" may be a key issue here, in that if evolution is broadly true, as most now assume, did we at some historic point become conscious and were not so before that point? The American psychologist Julian Jaynes has argued that consciousness has a real history, and that humans did not always consider themselves conscious: i.e. that consciousness does not automatically come with being *homo sapiens*. His argument, hugely simplified, is that after language was developed – say 60,000 years ago, but estimates vary widely – humans began to talk to themselves in their heads and this puzzled them. That novelty may have been related, Jaynes believes, to the Old Testament prophets dealing with this new phenomenon by claiming God was talking directly to them. Later, this self-conversation became an essential part of what we now call consciousness, which would imply that only humans have consciousness because only we have language. The Belgian scientist Cleeremans has argued that computers may *learn* to become conscious if and when they can endlessly talk to themselves about their plans and learn from doing this----which may be only an updated variant of Minsky's much earlier view of consciousness as arising from the need to review and adapt or debug our high-level plans of action. And Neil Lawrence has recently argued that this property derives from the fact that computers can communicate masses of coded data at high speed while we can only communicate slowly and with few slow symbols in a language. This, argues Lawrence, requires that to get information across in finite time, humans must share very large complex knowledge structures in their brains so that the language signal can be brief and inexplicit by calling upon these large shared structures, whose content we must also be able to review and (consciously) examine and adapt.

Are there moral questions about consciousness?

There has always been an assumption that if something were deemed conscious, then we could not ill-treat it or eat it, and so there would be immediate moral consequences for how we should treat machines if they achieved conscious status. We would no longer, in that case, be able to progress smoothly and guilt-free into the fully automated future where machines do all the work, because if they became conscious all the problems of living off slaves , or even the paid labour of conscious "others", would come back to haunt us.

Further Reading

Bello, P., and Bridewell, W., (2017) There is no agency without attention. *AI Magazine* 38(4)

Chalmers, D.J. (1996) *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Cleeremans A (2005) Computational correlates of consciousness. *Prog Brain Res.* 2005; 150:81-98.

Dennett, D.C. (1991) *Consciousness Explained*. Boston: Little, Brown, and Co.

Koch, C. (2019) *The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed*. Cambridge, MA: MIT Press.

Morch, H. (2017) Is matter conscious? http://nautil.us/issue/47/consciousness/is-matter-conscious?fbclid=IwAR3ypgDlLTidQSRcCxEl2OcFKNADrnu1bccOzHE-uIkAhLMegFHbsPzeLFo

Neumann, E. (1954) *The origins and history of consciousness.* Princeton, NJ: Princeton University Press

Parisi, Domenico. (2007). "Mental Robotics". In Chella & Manzotti (eds.), *Artificial Consciousness*. New York: Academic.

Strawson, G. (2005) 'The concept of consciousness in the Twentieth Century' (2016) in *Consciousness*, A. Simmons. (ed.) New York, NY: Oxford University Press.