www.theguardian.com /science/2012/oct/03/philosophy-artificial-intelligence

Philosophy will be the key that unlocks artificial intelligence

⋮ 03/10/2012

This article is more than **11 years old**

David Deutsch

AI is achievable, but it will take more than computer science and neuroscience to develop machines that think like people

To state that the human brain has capabilities that are, in some respects, far superior to those of all other known objects in the cosmos would be uncontroversial. The brain is the only kind of object capable of understanding that the cosmos is even there, or why there are infinitely many prime numbers, or that apples fall because of the curvature of space-time, or that obeying its own inborn instincts can be morally wrong, or that it itself exists. Nor are its unique abilities confined to such cerebral matters. The cold, physical fact is that it is the only kind of object that can propel itself into space and back without harm, or predict and prevent a meteor strike on itself, or cool objects to a billionth of a degree above absolute zero, or detect others of its kind across galactic distances.

But no brain on Earth is yet close to knowing what brains do in order to achieve any of that functionality. The enterprise of achieving it artificially – the field of "artificial general intelligence" or AGI – has made no progress whatever during the entire six decades of its existence.

Despite this long record of failure, AGI must be possible. That is because of a deep property of the laws of physics, namely the universality of computation. It entails that everything that the laws of physics require physical objects to do can, in principle, be emulated in arbitrarily fine detail by some program on a general-purpose computer, provided it is given enough time and memory.

So why has the field not progressed? In my view it is because, as an unknown sage once remarked, "it ain't what we don't know that causes trouble, it's what we know that just ain't so." I cannot think of any other significant field of knowledge where the prevailing wisdom, not only in society at large but among experts, is so beset with entrenched, overlapping, fundamental errors. Yet it has also been one of the most self-confident fields in prophesying that it will soon achieve the ultimate breakthrough.

In 1950, Alan Turing expected that by the year 2000, "one will be able to speak of machines thinking without expecting to be contradicted." In 1968, Arthur C Clarke expected it by 2001. Yet today, in 2012, no one is any better at programming an AGI than Turing himself would have been.

This does not surprise the dwindling band of opponents of the very possibility of AGI. But the other camp (the AGI-imminent one) recognises that this history of failure cries out to be explained – or, at least, to be rationalised away.

The very term "AGI" is an example of one such rationalisation, for the field used to be called "AI" – artificial intelligence. But AI was gradually appropriated to describe all sorts of unrelated computer programs such as game players, search engines and chatbots, until the G for "general" was added to make it possible to refer to the real thing again, but now with the implication that an AGI is just a smarter species of chatbot.

Another class of rationalisation runs along the general lines of: AGI isn't that great anyway; existing software is already as smart or smarter, but in a non-human way, and we are too vain or too culturally biased to give it due credit. This gets some traction because it invokes the persistently popular irrationality of cultural relativism, and also the related trope: "we humans pride ourselves on being the paragon of animals, but that pride is misplaced because they, too, have language, tools … And self-awareness." Remember the significance attributed to the computer system in the Terminator films, Skynet, becoming "self-aware"?

That's just another philosophical misconception, sufficient in itself to block any viable approach to AGI. The fact is that present-day software developers could straightforwardly program a computer to have "self-awareness" in the behavioural sense – for example, to pass the "mirror test" of being able to use a mirror to infer facts about itself – if they wanted to. As far as I am aware, no one has done so, presumably because it is a fairly useless ability as well as a trivial one.

Perhaps the reason self-awareness has its undeserved reputation for being connected with AGI is that, thanks to Gödel's theorem and various controversies in formal logic in the 20th century, self-reference of any kind has acquired a reputation for woo-woo mystery. And so has consciousness. And for consciousness we have the problem of ambiguous terminology again: the term has a huge range of meanings. At one end of the scale there is the philosophical problem of the nature of subjective sensations ("qualia"), which is intimately connected with the problem of AGI; but at the other end, "consciousness" is simply what we lose when we are put under general anaesthetic. Many animals certainly have that.

AGIs will indeed be capable of self-awareness – but that is because they will be General: they will be capable of awareness of every kind of deep and subtle thing, including their own selves. That does not mean that apes who pass the mirror test have any hint of the attributes of "general intelligence" of which AGI would be an artificial version. Indeed, Richard Byrne's wonderful research into gorilla memes has revealed how apes are able to learn useful behaviours from each other without ever understanding what they are for: the explanation of how ape cognition works really is behaviouristic.

Ironically, that group of rationalisations (AGI has already been done/is trivial/exists in apes/is a cultural conceit) are mirror images of arguments that originated in the AGI-is-impossible camp. For every argument of the form "you can't do AGI because you'll never be able to program the human soul, because it's supernatural," the AGI-is-easy camp has the rationalisation: "if you think that human cognition is qualitatively different from that of apes, you must believe in a supernatural soul."

"Anything we don't yet know how to program is called 'human intelligence'," is another such rationalisation. It is the mirror image of the argument advanced by the philosopher John Searle (from the "impossible" camp), who has pointed out that before computers existed, steam engines and later telegraph systems were used as metaphors for how the human mind must work. He argues that the hope that AGI is possible rests on a similarly insubstantial metaphor, namely that the mind is "essentially" a computer program. But that's not a metaphor: the universality of computation follows from the known laws of physics.

Some have suggested that the brain uses quantum computation, or even hyper-quantum computation relying on as-yet-unknown physics beyond quantum theory, and that this explains the failure to create AGI on existing computers. Explaining why I, and most researchers in the quantum theory of computation, disagree that that is a plausible source of the human brain's unique functionality is beyond the scope of this article.

That AGIs are "people" has been implicit in the very concept from the outset. If there were a program that lacked even a single cognitive ability that is characteristic of people, then by definition it would not qualify as an AGI; using non-cognitive attributes (such as percentage carbon content) to define personhood would be racist, favouring organic brains over silicon brains. But the fact that the ability to create new explanations is the unique, morally and intellectually significant functionality of "people" (humans and AGIs), and that they achieve this functionality by conjecture and criticism, changes everything.

Currently, personhood is often treated symbolically rather than factually – as an honorific, a promise to pretend that an entity (an ape, a foetus, a corporation) is a person in order to achieve some philosophical or practical aim. This isn't good. Never mind the terminology; change it if you like, and there are indeed reasons for treating various entities with respect, protecting them from harm and so on. All the same, the distinction between actual people, defined by that objective criterion, and other entities, has enormous moral and practical significance, and is going to become vital to the functioning of a civilisation that includes AGIs.

For example, the mere fact that it is not the computer but the running program that is a person raises unsolved philosophical problems that will become practical, political controversies as soon as AGIs exist – because once an AGI program is running in a computer, depriving it of that computer would be murder (or at least false imprisonment or slavery, as the case may be), just like depriving a human mind of its

body. But unlike a human body, an AGI program can be copied into multiple computers at the touch of a button. Are those programs, while they are still executing identical steps (ie before they have become differentiated due to random choices or different experiences), the same person or many different people? Do they get one vote, or many? Is deleting one of them murder, or a minor assault? And if some rogue programmer, perhaps illegally, creates billions of different AGI people, either on one computer or on many, what happens next? They are still people, with rights. Do they all get the vote?

Furthermore, in regard to AGIs, like any other entities with creativity, we have to forget almost all existing connotations of the word "programming". Treating AGIs like any other computer programs would constitute brainwashing, slavery and tyranny. And cruelty to children too, because "programming" an already-running AGI, unlike all other programming, constitutes education. And it constitutes debate, moral as well as factual. Ignoring the rights and personhood of AGIs would not only be the epitome of evil, but a recipe for disaster too: creative beings cannot be enslaved forever.

Some people are wondering whether we should welcome our new robot overlords and/or how we can rig their programming to make them constitutionally unable to harm humans (as in Asimov's "three laws of robotics"), and/or prevent them from acquiring the theory that the universe should be converted into paperclips. That's not the problem. It has always been the case that a single exceptionally creative person can be thousands of times as productive, economically, intellectually, or whatever, as most people; and that such a person, turning their powers to evil instead of good, can do enormous harm.

These phenomena have nothing to do with AGIs. The battle between good and evil ideas is as old as our species and will continue regardless of the hardware on which it is running. The issue is: we want the intelligences with (morally) good ideas always to defeat the evil intelligences, biological and artificial; but we are fallible, and our own conception of "good" needs continual improvement. How should society be organised so as to promote that improvement? "Enslave all intelligence" would be a catastrophically wrong answer, and "enslave all intelligence that doesn't look like us" would not be much better.

One implication is that we must stop regarding education (of humans or AGIs alike) as instruction – as a means of transmitting existing knowledge unaltered, and causing existing values to be enacted obediently. As Karl Popper wrote (in the context of scientific discovery, but it applies equally to the programming of AGIs and the education of children): "there is no such thing as instruction from without … We do not discover new facts or new effects by copying them, or by inferring them inductively from observation, or by any other method of instruction by the environment. We use, rather, the method of trial and the elimination of error." That is to say, conjecture and criticism. Learning must be something that newly created intelligences do, and control, for themselves.

I am not highlighting all these philosophical issues because I fear that AGIs will be invented before we have developed the philosophical sophistication to understand them and to integrate them into civilisation. It is for almost the opposite reason: I am convinced that the whole problem of developing AGIs is a matter of philosophy, not computer science or neurophysiology, and that the philosophical progress that will be essential to their future integration is also a prerequisite for developing them in the first place.

The lack of progress in AGI is due to a severe log jam of misconceptions. Without Popperian epistemology, one cannot even begin to guess what detailed functionality must be achieved to make an

AGI. And Popperian epistemology is not widely known, let alone understood well enough to be applied. Thinking of an AGI as a machine for translating experiences, rewards and punishments into ideas (or worse, just into behaviours) is like trying to cure infectious diseases by balancing bodily humours: futile because it is rooted in an archaic and wildly mistaken world view.

Without understanding that the functionality of an AGI is qualitatively different from that of any other kind of computer program, one is working in an entirely different field. If one works towards programs whose "thinking" is constitutionally incapable of violating predetermined constraints, one is trying to engineer away the defining attribute of an intelligent being, of a person: namely creativity.

Clearing this log jam will not, by itself, provide the answer. Yet the answer, conceived in those terms, cannot be all that difficult. For yet another consequence of understanding that the target ability is qualitatively different is that, since humans have it and apes do not, the information for how to achieve it must be encoded in the relatively tiny number of differences between the DNA of humans and that of chimpanzees. So in one respect I can agree with the AGI-is-imminent camp: it is plausible that just a single idea stands between us and the breakthrough. But it will have to be one of the best ideas ever.

Prof David Deutsch is a physicist at Oxford University and a pioneer of quantum computation. This is an abridged version of an essay that appears in the digital-only magazine Aeon