First published in the TLS 15 May 2020

**Computers don't give a damn**[1]

Tim Crane

In 1965 Herbert Simon, one of the founders of the new science of Artificial Intelligence (AI), wrote that "machines will be capable, within twenty years, of doing any work that a man can do". He was wrong of course — but maybe his mistake was only a matter of timing.

If Simon were to see what computing machines were capable of, fifty-five years after he made this remark, surely even he would be amazed. A single smartphone contains more computing power than all the world's computers in 1965 put together. And many of the philosophical arguments against the possibility of AI from the 60s and 70s fell flat on their face as the technology advanced. 1965 was also the year when the philosopher Hubert Dreyfus claimed that "no chess program can play even amateur chess" — true at the time, but proved false soon after. When in 1997 the IBM programme Deep Blue beat chess champion Garry Kasparov, this conclusively destroyed the idea that world-class chess was something computers can't do; Kasparov has commented recently that "today you can buy a chess engine for your laptop that will beat Deep Blue quite easily". And the familiar claim that computers could never really use their stored knowledge as well as human beings was shaken in 2011 by IBM's Watson programme, which won the top $1M prize on the American game show *Jeopardy*, beating the best human competitors.

AI sceptics used to claim that computers would never be able to recognise human faces or human speech, translate speech into text, or to convert handwriting to printed text. But today's phones can do all these things. Dreyfus had made gentle fun of the grand claims of AI, quoting a fanciful newspaper report from 1968 about "a new idea in gifts. . . . a genuine (if small) computer, that costs around $20. Battery operated, it looks like a portable typewriter. But it can be programmed like any big computer to translate foreign languages, diagnose illnesses, even provide a weather forecast". What seemed then like wild science fiction is now our everyday reality.

So Simon's claim may have been proved false, but maybe he was only a few decades out. The achievements of actual AI — that is, the kind of technology that makes your smartphone work — are incredible. These achievements have been made possible partly by developments in hard-

---

[1] Review of Brian Cantwell Smith, *The Promise of Artificial Intelligence*, MIT Press 2019

ware (in particular the increased speed and miniaturisation of microprocessors) and partly because of the access to incredible amounts of data on the internet — both factors that neither Simon nor Dreyfus could have predicted. But it means that enthusiastic predictions for AI are still popular. Many believe that AI can produce not just the "smart" devices that already dominate our lives, but genuine thinking machines. No one says that such machines already exist, of course, but many philosophers and scientists claim that they are on the horizon.

To get there requires creating what researchers call "Artificial General Intelligence" (AGI). As opposed to a special-purpose capacity — like Deep Blue's capacity to play chess — AGI is the general capacity to apply intelligence to an unlimited range of problems in the real world: something like the kind of intelligence we have. The philosopher David Chalmers has confidently claimed that "artificial general intelligence is possible … There are a lot of mountains we need to climb before we get to human-level AGI. That said, I think it's going to be possible eventually, say in the 40-to-100-year time frame". Philosophers John Basl and Eric Schwitzgebel are even more optimistic, claiming that it is "likely that we will soon have AI approximately as cognitively sophisticated as mice or dogs".

The intellectual enthusiasm for the possibility of AGI is matched by the vast sums invested in trying to make it a reality. In July 2019, Microsoft announced that it would invest $1 billion in Sam Altman's OpenAI, a for-profit company which aims to use AI for the "benefit of mankind as a whole". The British company Deep Mind, led by the computer scientist/neuroscientist Demis Hassabis, was bought by Google for $500 million in 2014. Deep Mind's best known achievement to date is the machine AlphaGo, which in 2016 beat Lee Sedol, the world champion of the ancient game of Go. Go is vastly more complex than chess — it sometimes said to be the most complex game ever created — and standard AI chess-playing methods had never been successfully applied to it. The computing methods used by AlphaGo are often touted as one of keys to "solving intelligence", as Deep Mind's own publicity puts it.

Brian Cantwell Smith's new book is provocative expression of scepticism about these recent claims on behalf of AI, from a distinguished practitioner in the field. His overall argument is based on a distinction between what he calls "reckoning" and "judgement". Reckoning is understood here in its original etymological sense: as calculation, like addition and subtraction. Judgement, by contrast, is described by Smith as "an overarching, systemic capacity or commitment, involving the whole commitment of the whole system to the whole world". Our thinking involves not

just some kind of simple on-off representation of things around us, but an entire emotional and value-laden involvement with the world itself. Computers have none of this. As the philosopher John Haugeland (a major influence on Smith) used to say, “computers don’t give a damn”. Giving a damn is a precondition of “judgement” in Smith’s sense, and anything that amounted to a real AGI would need to exercise judgement, and not simply calculate.

*The Promise of Artificial Intelligence* gives a brief and intelligible survey of two main stages in the history of AI. The first stage, starting in the 1960s, was what Haugeland christened “Good Old-Fashioned AI” (GOFAI) which solved computing problems by using explicit representations of general principles and applying them to particular situations. (Think of doing a mathematical proof or presenting an argument in logic.) “Second wave” AI, which started to emerge in the 1980s, began from the opposite end, so to speak: deriving general conclusions from huge amount of simple data as input. This kind of approach, variously called machine learning or deep learning, has had considerable success at things that GOFAI was very bad at, like pattern recognition, or updating knowledge on the basis of input.

First wave AI, it was often said, misconceived the nature of thinking: very little thinking resembles calculating or proving theorems. But Smith goes further, and argues that “the deeper problem is that it misconceived the world”. GOFAI assumed that “the world comes chopped up into discrete objects”, and because of this it analysed reasoning into its components by using formal logic (the basic ideas of which underlie modern computing). Smith argues that first wave theorising assumed that the world must be structured in the way that logic structures language: objects correspond to names, properties correspond to predicates or general terms. Things fit together in the world as symbols fit together in a logical language. Smith claims that this is one main reason why the GOFAI project failed: it failed to take account of the “fabulously rich and messy world we inhabit”, which does not come in a “pre-given” form, divided up into objects.

Second wave AI, according to Smith, does not make this mistake. It does not assume a “pre-given” ontology or structure for the world, and for that reason, he argues, it has made progress in the areas where GOFAI failed: in particular, with tasks like face recognition, text processing and (most famously) the game of Go. The distinctive feature of deep learning machines is their ability to detect patterns in large (sometimes huge) amounts of data. The machines “learn” by being given an indication by the programmer which patterns are the important ones, and after a while they can produce results (e.g. moves in a game) which surprise even the programmers. This

is in contrast to first wave AI programmes which attempted to anticipate in advance how input from the real world should be responded to in every conceivable situation. Those early AI machines that worked only did so in very constrained made-up environments, sometimes called “microworlds”.

Nonetheless, Smith thinks that we should not be too optimistic about the ability of second wave AI to create AGI. Machine learning may not start with general rules which make ontological assumptions, but it does start with data that is already processed by humans (e.g. things that we classify as faces, or as road traffic at an intersection and so on). Much machine learning, as Smith says, is “dedicated to sorting inputs into categories of manifest human origin and utility”. So even if they are more sensitive to the messy world, second wave AI machines are still tied up with the programmers’ own classifications of reality — indeed, it’s hard to see how they could be otherwise designed.

Smith is surely right that AI’s recent successes give us little or no reason to believe in the real possibility of genuine thinking machines. His distinction between reckoning and judgement is an important attempt to identify what it is that is missing in AI models. In many ways (despite his protest to the contrary) it echoes the criticisms of Dreyfus and others, that AI will not succeed in creating genuine thinking unless it can in some way capture “common sense”. And just as common sense (part of Smith’s “judgement”) cannot be captured in terms of the “rules and representations” of GOFAI, nor can it be captured by massively parallel computing drawing patterns from data.

To make this point about judgement, Smith does not actually need the more ambitious ontological claims, that the world does not have natural divisions or boundaries, that all classification is simply a result of human interest, and so on. Maybe these claims are true, maybe not — many centuries of philosophy has wrestled with them, and they are worth debating. But we should not need to debate them in order to identify the fundamental implausibility of the idea that AGI is on the horizon.

This implausibility derives from something which is intrinsic to the success of AI itself. For despite the sophistication of machine learning, the fact remains that like chess, Go is still a *game*. It has rules and a clear outcome which is the target for players. Deep learning machines are still being used to achieve a well-defined goal — winning the game — the meaning of which can be articulated in advance of turning on the machine. The same is true of speech and face recognition software. There is a clear goal or target — recognising the words and faces —and successes and

failures in meeting this goal are the input which helps train the machine. (As Smith says, “recognition” here means: correctly mapping an image onto a label: nothing more than that.)

But what might be the goal of “general intelligence”? How can we characterise in abstract terms the problems that general intelligence is trying to solve? I think it’s fair to say that no-one — in AI, or philosophy, or psychology — has any idea how to answer this question. Arguably, this is not because it is an incredibly difficult empirical question, but rather that it is not obviously a sensible one. I suppose someone might say, in the spirit of Herbert Simon (whose famous AI programme was called the “General Problem Solver”), that general intelligence is the general ability to solve cognitive problems. This might seem fine until we ask ourselves how we should identify, in general terms, the cognitive problems which we use our intelligence to solve. How can we say, in general terms, what these problems are?

Consider for example, the challenges faced in trying to create a genuine conversation with a computer. Voice assistants like Siri and Alexa do amazingly well in “recognising” speech and synthesising speech in response. But you very quickly get to the bottom of their resources and reach a “canned” response (“here are some webpages relating to your inquiry”). One reason for this, surely, is that conversation is not an activity that has one easily expressible goal, and so the task for the Siri/Alexa programme cannot be specified in advance. If the goal of conversation were to find information about a subject-matter, then directing to you to a website with relevant information could be one reliable way of achieving that goal. But of course this is not the sole thing to which we direct our intelligence when talking with others.

What, then, is the overall goal of conversation? There isn’t one. We talk to pass the time, to express our emotions, feelings and desires, to find out more about others, to have fun, to be polite, to educate others, to make money… and so on. But if there is no single goal of conversation, then it is even less likely it is that there is one goal of “general intelligence”. So no wonder AI researchers struggle to even define the “task domain” for AGI.

As Smith’s book shows, the claims for the possibility of AGI ignore the huge differences between the relatively well-defined areas where AI has succeeded, and the barely-defined domain of “general intelligence”. This is, on its own, enough of a reason for scepticism about extrapolating beyond the successes of actual AI to the real possibility of AGI. Smith’s arguments about the ontological assumptions of AI, whatever their merits, are not necessary to make this point.

Yet I suspect that many still have this lingering sense that AGI must be possible, and that the difference between real human thinking and what computers do is just a matter of complexity. What lies behind this conviction? One widespread idea is that since the human brain is just a complex biological (and therefore material) machine, it must be possible in principle to artificially reproduce what the human brain does (thinking, perceiving, feeling, imagining, being conscious etc) by building something that functions in exactly the same way as the brain. And whatever we thereby build will be an artificial version of our mental processes: an AGI.

This argument is based on two ideas: first, that thinking and other mental processes go on in the brain; second, the brain is a machine or mechanism. So if we can uncover the principles that make this mechanism work, and we have adequate technology — the argument goes — then we should be able to build a machine that implements these principles, without leaving anything out. One of the pioneers of deep learning, Yoshua Benigo, puts it this way: “I don’t know how much time it’s going to take, but the human brain is a machine. It’s a very complex one and we don’t fully understand it, but there’s no reason to believe that we won’t be able to figure out those principles”.

Of course building an artificial copy of a real brain is nowhere close to today’s scientific reality. But if we believe that we are at bottom material beings — if we take away all our matter then there is nothing left of us — then such replication seems possible in principle, even if it is never actually realised. Suppose, then, a brilliant scientist of the future could replicate in an artificial construction everything a person (and its brain) does. Obviously this replica would also be able to think, since thinking is one of the things a person does. It is undeniable that making an artificial replica of a person and all its features would be making an “artificial intelligence” in an an obvious sense: simply because intelligence is one of the features of people.

The question is, what does this have to do with AI? If the way to create a real artificial thinker is to find out first how the brain works, then you would expect AI researchers to try and figure out how the brain or the mind actually works — that is, to become neuroscientists or psychologists. But this is not how AI researchers operate. Just as the invention of the aeroplane did not require building something that flies in exactly the way a bird flies, so the inventors of AI did not feel bound to copy the actual workings of human brains. And despite the fact that deep learning computers use what are called “neural networks”, the similarity to the brain here is at a very abstract level. Indeed, many pioneers of deep learning occupy themselves with very abstract questions

about general intelligence — Benigo says his goal is “to understand the general principles of intelligence” — rather than with the messy business of the actual working of the human mind or brain.

This lack of focus on the way human minds (or brains) actually function goes back to the beginnings of AI, and it was clearly one of AI’s strengths. By ignoring the complexity of actual human thinking, and the messy “wetware” of the actual human brain, AI could get machines to solve difficult problems without having to bother with how we would solve them. Watson, the IBM website tells us, “is not bound by volume or memory. Watson can read millions of unstructured documents in seconds”. That’s not something we can do. The learning involved in training AlphaGo involved millions of practice games of Go — as cognitive scientist and deep learning sceptic Gary Marcus has pointed out, this is “far more than a human would require to become world class at Go”. This means that whatever it explains their success, it is not the similarity of these machines to human thinkers. So IBM should not claim that Watson is “built to mirror the same learning process that we have, a common cognitive framework that humans use to inform their decisions”.

As it actually now is, AI does not rely substantially on any detailed psychological or neuroscientific research. So the fact that the brain is a material mechanism which could, in principle, be replicated artificially gives no support to the idea that AI *as it actually is* could build an AGI. In fact, given the way AI has actually proceeded, in splendid isolation from neuroscience, it is likely that any attempt to replicate the brain would require ideas very different from those traditionally used by AI. To say this is not to disparage AI and its achievements, it is just to emphasise the obvious fact that it is not, and has never been, a theory of human thinking.

Philosophical and scientific discussions of AI have tended towards one of two extremes: either that genuine artificial thinking machines are just on the horizon, or they are absolutely impossible in principle. Neither approach is quite right. On the one hand, as Smith, Marcus and others have explained, we should be sceptical that recent advances in AI give any support to the real possibility of AGI. But on the other, it is hard to deny the abstract philosophical claim that if you could replicate a human brain in such a way that would produce something that did everything the brain did, then that thing would be a “thinking machine” in one clear sense of that phrase. However, this mere possibility does not mean that today’s AI is anywhere near creating genuine thinking machines. On the contrary: when you examine the possibility more closely, it shows why AI is unlikely ever to do this.