GRESHAM COLLEGE

14 MAY 2019

IS THERE DANGER AHEAD WITH AI?

PROFESSOR YORICK WILKS

The idea of an artificial superintelligence goes back at least to I.J. Good's "ultraintelligent machine" (1965) and Turing's (1951) warning: "If a machine can think, it might think more intelligently than we do, and then where should we be? Even if we could keep the machines in a subservient position, for instance by turning off the power at strategic moments, we should, as a species, feel greatly humbled. This new danger is certainly something which can give us anxiety." It has also been a staple of science fiction, most famously HAL9000, the malign spaceship computer in Kubrick's film of Arthur C. Clarke' story *2001*. Science fiction is relevant here because I will argue that science fiction has been a motivation in the revival of the notion in Bostrom's best-selling book *Superintelligence* (2014) which has been applauded by Musk, Gates, and Hawking among others, all of whom to share something of Bostrom's pessimistic view about AI's futures.

In a recent issue of *Wired*, the same idea lay behind an article that toyed with the timely fantasy of an AI US President (Davis, 2017), where the magazine declared that it would be driven by machine learning and therefore erratic and inscrutable as to its reasons for action. Confusingly, the article also claimed that the artificial President would have to aim for the greatest good of the greatest number, the standard ethical calculus known as utilitarianism or consequentialism, and usually contrasted with an ethical doctrine based on principles and rules, conventionally attributed to the philosopher Kant. But there is no reason at all for that assumption; one could just as well program a political computer to obey clear rules such as "thou shalt not steal". This will be important later because these ethical considerations are right at the centre of problem of how a future superintelligence might treat us. Very recently, and by chance from the same Oxford Institute for the Future that houses Bostrom, some recent work by Crockett and others claims that surveys show that people tend to prefer those who they perceive to work with ethical rules (Everett et al., 2017) rather than by calculating consequences. People might therefore also prefer to deal with intelligent computers that had clear ethical rules rather than calculating consequences. Film buffs will remember that it was such calculations in *2001* that led HAL9000 to the conclusion that it had to destroy the human crew of its ship and itself.

Bostrom's *Superintelligence* (SI) is a wide-ranging essay (2016) that has raised important questions about the future of intelligent machines and their possible malign developments. But, and perhaps surprisingly, it is not about technical developments in AI nor a philosophical analysis of the concept of SI, but much more akin to science fiction. When you encounter one more scary-AI article, designed to terrify you with possibilities just around the corner that may end the civilization you know and love, the first thing to ask is whether the author knows anything about the technology, and possibly whether they have any hands-on experience of it. If not, they may be in no better a seer than scifi pioneers H G Wells or Jules Verne, who were often wrong, though sometimes right.

None of this implies that philosophers should not comment on the future course of a science: long ago Dreyfus (1992) and Dennett (1987) questioned the basis of AI methodology, and both did the subject a great deal of good. That is a very different activity from the claim to have analysed AI and its potential and to declare them socially

dangerous, which is what Bostrom is doing, but without any detailed diagnosis of the methods, goals and successes of AI. Weizenbaum (1976) warned us of the danger of his own ELIZA program, but at least he built it first.

What then is Superintelligence?

Bostrom offers three attributes of superintelligences after a preliminary definition that he concedes is vague: "intellects that greatly outperform the best current human minds across many very general cognitive domains". The "many domains" are needed in the definition because we have long had examples of computing that outperform us in narrow domains from arithmetic to chess. Those attributes are speed, collectivity and quality.

Speed alone cannot be sufficient for superintelligence because very fast arithmetical calculators have existed for decades and no one considers them intelligent, so the question is how far speed is necessary for SI? It could certainly be useful: superfast stock-trading algorithms have revolutionised the stock market and made fortunes, though a superfast fast thinking teacher or diplomat would probably not be a great success.

But Bostrom argues that, to a very fast mind, "events in the external world would appear to unfold in slow motion" (p.64). Yet I can imagine flipping through a phone directory using my superfast mind to compute statistics about the owners of phones and their names and addresses. But that would not speed up the movements of my hands which would be normal. In his story *The New Accelerator*, H.G. Wells' hero on the eponymous drug passes through rooms unnoticed at very high speed where his friends appear always stationary to him. Perhaps Bostrom above is just slipping from having a fast mind to a fast body like Wells' hero?

As to the possible connections of SI to realistic speed in a computing application, no discussion would be complete without considering the speed/storage trade-off. A SI might be so because it had a very large memory capacity with a very efficient and reliable indexing and retrieval mechanism. The requirement for pattern recognition is the latter, not raw speed, which can always be ramped up with enough parallelism, so it is hard to relate speed as such to any advance in the intelligence of a program

Arguments about human intelligence, collective and individual, are important to Bostrom because he sees their historical augmentation as steps towards superintelligence itself. Yet this is not obvious at all since there is no general consensus that human intelligence does increase over time. The so-called Flynn effect (1987) claims to show that IQ tests have to be constantly recalibrated because people everywhere are getting so much better at them, and showing IQ rises of several percent a decade. But this is widely disputed, and not only by reactionaries who believe that all human thought has only been footnotes to Plato and Aristotle. There are good reasons to believe the Flynn effect is simply an artefact of testing, and there are many who believe IQ is in fact falling, at least in the West, on the standard eugenic grounds of its partial heritability and the falling fertility of the best educated (and thus intelligent) women. We do not need to have a view on that issue here but the truth or falsehood of the Flynn effect certainly bears on the likelihood of SI emerging by biological or natural means.

Bostrom concludes: "Superintelligence in any of these [three] forms could, over time, develop the technology necessary to create any of the others" (p. 79). Yet it is an odd conclusion that, say, vastly speeded up intelligence would lead to leaps in quality, or vice versa. Why would one imagine that, given their independence? Processing speed has made great gains but those have not of themselves led to any improvement in AI quality, which has come from better theories and algorithms and from increased capacity and processor size. This last has enabled computations that were envisaged in the past but not carried out, e.g. on large matrices, or the success of "deep learning" methods (e.g. Bengio et al., 2015) in vision. But these advances have not been dependent on processor speed as such. Nor does our own intelligence in any way depend on speed, since brain processes seem if anything characterised by their extraordinary *slowness*, in electronic terms, As he concedes: "the human brain…is incapable of rapidly performing any computation that requires a large number of sequential operations" (p. 72).

Achieving Superintelligence

Having, as he believes, defined SI, Bostrom asks how soon it may come about. Like almost everything else in this work, it comes in three forms: slow, fast and moderate, and he believes the rate will be governed by the equation:

$$dI/dt = O/R$$

which gives us the rate of change of Intelligence (on the left) as the optimization power O, divided by the *recalcitrance* R. It soon becomes clear that this is decorative rather than quantitative mathematics and is not meant to be taken seriously. The equation itself is not wholly serious because neither O nor R seems to be an integrable function of time. It is metaphorical math whose key notion is R: which is what is holding a system's optimization back. The equation's feel and shape is very much that of Ohm's law:

$$I = V/R \quad \text{i.e. Current = Voltage/Resistance}$$

which conveniently uses two of the same letters and may perhaps have been its inspiration. The key notion of recalcitrance is explained by the rather good metaphor of solving a jigsaw puzzle, which initially goes quickly as you get the corners and edges (low recalcitrance) and quickly at the end with few pieces left to fill, but which is much harder between those phases, when there is then high recalcitrance.

In describing recalcitrance as a hindrance to SI, Bostrom also implicates content size, by which he means the quantity of knowledge that the SI has. Here he falls back for illustration on IBM's WATSON [Ferruci, 2010] and the range of textual knowledge it searches in answering questions. He considers extending a WATSON to contain, say, the Library of Congress, after which: " the system knows much more than any human being and thinks vastly faster; it has become (at least) weakly superintelligent." (p87).

This is an interesting example and a welcome element of concreteness in a very speculative discussion, but it is also revealing in another way. Whatever an SI is to be, it would surely have much of the kind of textual knowledge we all have, and most human knowledge is still, as they say, in unstructured form, i.e. texts in libraries and so on. An SI might well have more of it than any human and so quantity of text content is a relevant consideration. But WATSON has none of the features one might hope for in an SI: its success in the *Jeopardy* competition rested largely on the fact that it did not have a "human equation": the human delay in button pushing. It could always push the button ahead of its human competitors because it did not have brain-to-arm nerve transmission. More seriously, WATSON represents the optimization of a particular set of 1990s information extraction and question answering techniques over a fixed set of documents. It is not advanced AI at all but a set very shallow techniques honed over a decade, and was never more than state of the art. It is certainly important work, but it has none of what one would demand of any SI such as understanding the texts it processed. That is the key intellectual problem underlying the AI developments one can conveniently capture with the term "semantic web", the project that asked how a web of human documents can understand its own content (as the current world wide web certainly does not) and what kind of representation of the documents that would require? [Wilks, 2008].

The key term "representation" has entered and points to the fundamental dispute going on within AI as to the kind of representation of knowledge that is required, if any, for tasks like really understanding language: must it be expressed in symbols as humans do on paper, or is it some statistical, mathematical or other quantitative representation? I stress this because WATSON has no such representation, or even complex quantitative forms like vectors, expressing the content of the texts it searches. It is a very simple machine indeed, but no hint of that appears in Bostrom's discussion because he is not really concerned with the real research content of AI, or even SI, at all.

How intelligent will an SI be?

One can see Bostrom's SI as a late response to what Papert (according to Minsky, 1974) called the "superhuman-human fallacy" in his definition of the goal of AI, which was not to model superhuman functions, but just those of normal human activity, which he thought were the interesting and demanding problems. Bostrom's whole position is that, on the contrary, AI can and must lead onward to the SI, which Papert explicitly rejected.

He discusses what actual faculties an SI should have, beyond them being many and varied, all superior to our own. He proposes concrete abilities (p.119) such as solving the "peptide folding problem, and by (p.122) "sending out large numbers of von Neumann probes" to locate and report on habitable planets and return at half the speed of light. These are no doubt real problems, but most readers will not understand the issues involved in their construction and the constraints on their solutions. And does one believe for a moment that even Bostrom has the skills to understand the problems in the two very different scientific fields—possibly no one fully understands both—so how can he expect us to accept that he knows enough to create and manipulate these examples to show that they plainly require an SI for their solution? This makes them implausible examples for getting across the basic capacities of an SI.

The crucial question is what goals would an SI have, and what would it want to do or achieve? Bostrom has an argument that it would differ from us because its goals would be fixed rather than mutable like ours. SIs would have permanent goals, though it is quite unclear why. Why might they not change their final goals as we do with changing circumstances, not all of which they can, presumably, control? Of course, we can program any automaton to have fixed goals but that is irrelevant here since, by definition, a Bostrom SI can override and control any such pre-programming by us, so the fixed goals he postulates for it must be its own choice. By why should it so limit itself? He does not tell us.

It is worth noting how utterly disjoint all Bostrom's thinking is from any kind of evolutionary adaptation, the core process that has produced us as a species. He mentions evolutionary adaptation once or twice but only to dismiss it as not likely to produce desirable results. Yet our world is one in which the idea of fixed, preset, goals of the kind he envisaged for an SI would be, almost by definition, a death sentence for any life-form having them, especially if it passed those same fixed goals on to any SI it created.

He argues strongly (pp. 136-7) than an SI will be a singleton: that there will be just one of them. And again, one can ask, why and how could he know that? Many might be created, so this implies they would be in competition and the singleton SI would emerge after destroying the others. But why should they not cooperate and share: since there is good reason to think enhanced cooperation is a mark of intelligence, especially with equals. The SI would also, presumably, have to remain a singleton and renounce the ability to create others lest they turn on their parent later, thus again frustrating any possible future evolution. Science fiction at this point seems to veer off into an imaginary anthropology of parent killing by SI offspring.

But none of this is very good science fiction, even though he assured us earlier that SIs will not have the human goal of carrying off girls in flimsy dresses like cartoon scifi monsters, though he promptly gives us several pictures of monsters and girls in those very dresses. (p.129). The style is important here: philosophical prose applied not to complex argument but to social speculation, along with a boyhood attraction to monsters and scifi.

The outcome of an SI will be an "existential catastrophe" for humans.

The "existential catastrophe", by which he means future human extinction provoked by an SI, is not defined nor is the phrase in the index, but his core prediction is that "while weak, an AI behaves cooperatively…. When sufficiently strong—without warning or provocation---it strikes, forming a singleton, and begins directly to optimize the world according to the criteria implied by its final values." (p. 144).

No science fiction is mentioned in the discussion of this doom scenario but it is hard not to see again behind this the notorious HAL 9000: HAL was a singleton, good at everything (it learned lip reading at superhuman speed) and had a fixed final goal on the mission quite different from, and unknown to, the humans passengers on its ship.

What Bostrom calls "malignant failure" by an SI seems to be intimately connected to notions of "decisive strategic advantage"—a term from business studies and moderately comprehensible ---and "infrastructure profusion" which is illustrated (p.15) by a long and tedious digression on an SI making paperclips. It makes too many paperclips, absorbing all the resources of the earth to make more and more of them Even if told to make only a fixed number it expends endless resources checking and rechecking that it has the exactly the right number, thus possibly checking for ever. None of this seems remotely plausible since the *Sorceror's Apprentice*, even if applied to the programming of a mini-capitalist's clip-making robot in a garage, let alone anything worthy of the name SI.

What then can be done to ward off catastrophe?

Bostrom's answers to this key question are surprisingly conventional and include putting the SI without arms in a box.

But his serious suggestions are motivational, to use his term: following Asimov's laws of robotics, one might install in the SI limited final goals, or try a set of goals out in a simpler, safer, system and then add in the full SI.

This last is complex and may be one of the best of Bostrom's ideas: He proposes various graded incentives, including reward keys hidden away in a safe place. Though, as hackers show us daily, what one person can hide another can find and why would an SI not be at least as good as us at locating the hidden goodies? None of these constraints give much comfort if one accepts the plausibility of the malignant or treacherous turn, and that is perhaps the real issue: the plausibility of all this, and whether humans are any good at estimating and dealing with huge problems or events that have a very low probability of occurring. All history suggests they are not, given their inability to deal with catastrophes with far higher probability such as global warming.

Bostrom finds Asimov's classic laws of robotics (1964) inadequate as motivational constraints, as all commentators do. An author he might have looked to here is Thomas Schelling (1969), whose work on complex game theoretic principles that might extend Asimov's laws take the form of constraints on what an entity MUST NOT DO. One might work back from these to plans of action and strategy in a way that would be harder to unpick for the SI than Asimov's over-simple rules. Considerations like this bring back the earlier preference we seem to have for people who work with moral rules rather than mere consequence calculations, and the question of whether that preference will transfer to an artificial actors we trust. Yet consequence calculation cannot be avoided in intelligent agents as the automated car case shows clearly: if an accident is unavoidable and some group must be hit and possibly killed, the car will have to decide which one to damage and why, and no "do not kill" rule will then apply.

No account is ever given by Bostrom of where an SI's final values, inconsistent with human ones, might come from, nor why they should be hostile to humans. Yet the whole thrust of his pessimistic argument, rests on assumptions about the nature of these values, and the inadequacy of anything like Asimov's laws to deal with them. It has been part of most human religions, speaking purely historically and not theologically, to worship the creator of humanity and this has happened without any particular assumptions about the intelligence of the creator, who was generally thought admirable and an object of positive emotions, even if omniscient and omnipotent. Why therefore might we not assume that an SI would feel similarly about its own creator, humanity, independently of that creator's IQ level, which it might well, realise to be lower than its own? This is surely at least as plausible as Bostrom's assumption to the contrary and at least has some historical-anthropological backing.

The Value-Loading Problem

The most vexed question for Bostrom's argument is what he calls the value-loading problem: as he puts it (p. 228): "If we cannot transfer human values into an AI by typing out full-blown representations in computer code, what else might we try?" The obvious overall answer is to instil values we approve of in the SI, values that will be favourable to us as humans and not destroy or enslave us. "How" he asks, "can we get an SI to do what I want?" (p.256). This is the key (unanswered) question in the work, and it the very same question Asimov sought to answer.

One concrete version of these issues, that both brings us back to our ethical starting point and has been the subject of much public discussion in connection with automated cars is Phillipa Foot's "trolley problem" (1978) as a way of analysing the kill-or-save choices an automated car will be required to make. This has been a simple but influential example of a fully described ethical problem to do with who and how many people must be killed plus an automated device that must make decisions. There are no such concrete ethical problems described for an SI, beyond a suggestion that we should perhaps defer to its values (p 258), because it knows more than we do *ex hypothesi*. This is a very odd conclusion to work that has argued throughout that an SI's basic values would have no connection to human ones and might well involve destroying us.

Bostrom derives a notion of "Controlled Extrapolated Volition" from Yudkowsky, an author who seems to have had much influence on his thinking:, and who argues, as mentioned above, that an SI should be put in charge of us since, in any case, humans cannot agree about the content of ethics. Yudkowsky is one of a range of writers on decision theory and economics who have influenced Bostrom's thinking on value loading. There is a substantial literature now on AI and ethics but none of it finds any place in this work. Earlier I mentioned Dreyfus' (1992) argument that AI could only succeed or have values if and when it had produced entities that grew up and learned values as we do. His position is more relevant now that machine learning has returned to AI's centre, though none of this is noticed by Bostrom, apart from a passing reference to "reinforcement" learning, a relatively simple approach within what is now a large and complex technical subject. Bostrom's arguments all rest on a view of AI with inference and (possibly malign) goals at its centre, even though this combination is far less central in the AI world of big data and machine learning.

None of this implies we should be sanguine about the future of AI or its potential for harm, but SI is highly unlikely to come about in the way Bostrom imagines. Edward Geist recently (2017) had some wise words about the real dangers: "AI-enhanced technologies might still be extremely dangerous due to their potential for amplifying human stupidity. The AIs of the foreseeable future need not think or create to sow mass unemployment, or enable new weapons technologies that undermine precarious strategic balances. Nor does artificial intelligence need to be smarter than humans to threaten our survival—all it needs to do is make the technologies behind familiar 20th-century existential threats faster, cheaper, and more deadly."

References

Asimov (1964 Introduction. *The Rest of the Robots*. New York: Doubleday.

Bengio, Y., LeCun, Y., Hinton, G. (2015). Deep Learning. *Nature*. **521**: 436–444

Bostrom, N. (2016) *Superintelligence.* Oxford: Oxford University Press.

Davis, J. (2017) Hear me out: let's elect an AI as President. *Wired Magazine*. **URI** https://www.wired.com/2017/05/hear-lets-elect-ai-president/

Dennett, D. (1987) *The Intentional Stance (6th printing), Cambridge, MA: The MIT Press.*

Dreyfus, H. (1992) *What Computers Still Can't Do: The Limits of Artificial Intelligence*. Cambridge, MA: The MIT Press.

Everett, J., Pizarro, D., and Crockett, M. (2017) Why are we reluctant to trust robots. *The Guardian*. ***URI*:** https://www.theguardian.com/science/head-quarters/2017/apr/24/why-are-we-reluctant-to-trust-robots

Ferrucci, D.; et al. (2010). Building Watson: An Overview of the DeepQA Project" *AI Magazine*. **31** (3).

Flynn J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*. **101** (2): 171–191.

Foot, P. (1978) *The Problem of Abortion and the Doctrine of the Double Effect* in *Virtues and Vices*. Oxford: Basil Blackwell.

Geist, E. (2017) (Automated) planning for tomorrow: Will artificial intelligence get smarter? *Bulletin of the Atomic Scientists* 73(2). ***URI*:** http://thebulletin.org/2017/march/automated-planning-tomorrow-will-artificial-intelligence-get-smarter10558

Good, I. J. (1965) Speculations Concerning the First Ultraintelligent Machine . In F. Alt & M. Ruminoff (eds.), *Advances in Computers,* volume 6. Academic Press, 31-88.

Minsky, M. (1974) *A Framework for representing knowledge.* ***URI*:** http://hdl.handle.net/1721.1/6089

Schelling, T. (1969) Models of segregation, *American Economic Review*, 1969, 59(2), 488–493

Turing, A. (1951) Can digital machines think? Lecture broadcast on BBC Third Programme; typescript at turingarchive.org.,

Weizenbaum, J. (1976) *Computer Power and Human Reason*. San Francisco: W.H. Freeman.

Wilks, Y. (2008), The Semantic Web: Apotheosis of annotation, but what are its semantics? *IEEE Intelligent Systems,* Volume 23 (3), 41 - 49.