THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Smart Machines are Not a Threat to Humanity

Citation for published version:
Bundy, A 2017, 'Smart Machines are Not a Threat to Humanity', *Communications of the ACM*, vol. 60, no. 2, pp. 40-42. https://doi.org/10.1145/2950042

Digital Object Identifier (DOI):
10.1145/2950042

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Communications of the ACM

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 12. May. 2024

Smart Machines are Not a Threat to Humanity *

Alan Bundy

June 15, 2016

Abstract

The root of my argument is that any AI threat comes, not from machines that are too smart, but from machines that are too dumb. Such dumb machines pose a threat to individual humans, but not to humanity. Worrying about machines that are too smart distracts us from the real and present threat from machines that are too dumb.

Concerns have recently been widely expressed that Artificial Intelligence presents a threat to humanity. For instance, Stephen Hawking is quoted in [Cellan-Jones, 2014] as saying:

> "The development of full artificial intelligence could spell the end of the human race."

Similar concerns have also been expressed by Elon Musk, Steve Wozniak and others.

Such concerns have a long history. John von Neumann is quoted by Stanislaw Ulam [Ulam, 1958] as the first to use the term *the singularity*[1] , i.e., the point at which artificial intelligence exceeds human intelligence. Ray Kurzweil has predicted that the singularity will occur around 2045 [Kurzweil, 2005] — a prediction based on Moore's Law as the time when machine speed and memory capacity will rival human capacity. I.J. Good has predicted that such super-intelligent machines will then build even more intelligent machines in an accelerating 'intelligence explosion' [Good, 1965]. The fear is that these super-intelligent machines will pose an existential threat to humanity, e.g., keep humans as pets or kill us all [Warwick, 2004] — or maybe humanity will just be a victim of evolution.

I think the concept of the singularity is ill-conceived. It is based on an oversimplified and false understanding of intelligence. Moore's Law will not inevitably lead to such a singularity. Progress in AI depends not just on speed and memory size, but also on developing new algorithms and the new concepts that underpin them. More crucially, the singularity is predicated on a linear model of intelligence, rather like IQ, on which each animal species has its place, and along which AI is gradually advancing. *Intelligence is not like this.* As Aaron Sloman, for instance, has successfully argued, intelligence must be modelled using a multi-dimensional space, with many different kinds of intelligence and with AI progressing in many different directions [Sloman, 1995].

AI systems occupy points in this multi-dimensional space that are unlike any animal species. In particular, their expertise tends to be very high in very narrow areas, but non-existent elsewhere. Consider, for instance, some of the most successful AI systems of the last few decades.

Deep Blue: was a chess-playing computer, developed by IBM, that defeated the then world champion, Garry Kasparov, in 1996. Deep Blue could play chess better than any human, but could not do anything other than play chess — it couldn't even move the pieces on a physical board.

Tartan Racing's Boss: was a self-driving car, built by Carnegie Mellon University and General Motors, which won the DARPA Urban Challenge in 2007. It was the first to show that self-driving cars could operate safely alongside humans, and so stimulated the current commercial interest in this technology. Tartan Racing couldn't play chess or do anything other than drive a car.

*Thanks to Stephan Schulz, Lucas Dixon, the St Andrews University Student Debating Society and two anonymous CACM reviewers for feedback on earlier versions.

[1] `https://en.wikipedia.org/wiki/Technological_singularity` (accessed 31^{st} December 2015)

Watson: also developed by IBM, was a question answering system that in 2011 beat the World champions at the Jeopardy general-knowledge quiz game. It can't play chess or drive a car. IBM are developing versions of Watson for a wide range of other domains, e.g., in healthcare, the pharmaceutical industry, publishing, biotechnology and a chatterbox for toys. Each of these applications will be similarly narrowly focused.

AlphaGo: was a Go-playing program, developed by Google's DeepMind, that beat the World champion, Lee Sedol, 4-1 in March 2016. AlphaGo was trained to play Go using deep learning. Like Deep Blue, it required a human to move the pieces on the physical board and couldn't do anything other than play Go, although DeepMind used similar techniques to build other board-game playing programs.

Is this situation likely to change in the foreseeable future? There is currently a revival of interest in *Artificial General Intelligence*, the attempt to build a machine that could successfully perform any intellectual task that a human being can. Is there any reason to believe that progress now will be faster than it has been since John McCarthy advocated it 60 years ago at the 1956 inaugural AI conference at Dartmouth? It's generally agreed that one of key enabling technologies will be commonsense reasoning. A recent CACM Review article [Davis & Marcus, 2015] argues that, while significant progress has been made in several areas of reasoning: temporal, geometric, multi-agent, etc., many intractable problems remain. Note also that, while successful systems, such as Watson and AlphaGo, have been applied to new areas, each of these applications is still narrow in scope. One could use a 'Big Switch' approach, to direct each task to the appropriate narrowly-scoped system, but this approach is generally regard as inadequate in not providing the integration of multiple cognitive processes routinely employed by humans.

I am not trying to argue that Artifical General Intelligence is, in principle, impossible. I don't believe that there is anything in human cognition that is beyond scientific understanding. With such an understanding will surely come the ability to emulate it artificially. But I'm not holding my breath. I've lived through too many AI hype cycles to expect the latest one to deliver something that previous cycles have failed to deliver. And I don't believe that now is the time to worry about a threat to humanity from smart machines, when there is a much more pressing problem to worry about.

That problem is that many humans tend to ascribe too much intelligence to narrowly focused AI systems. Any machine that can beat all humans at Go must surely be very intelligent, so by analogy with other world-class Go players, it must be pretty smart in other ways too, mustn't it? No! Such misconceptions lead to false expectations that such AI systems will work correctly in areas outside their narrow expertise. This can cause problems, e.g., a medical diagnosis system might recommend the wrong treatment when faced with a disease beyond its diagnostic ability, a self-driving car has already crashed when confronted by an unanticipated situation. Such erroneous behaviour by dumb machines certainly presents a threat to *individual humans*, but not to *humanity*. To counter it, AI systems need an internal model of their scope and limitations, so that they can recognise when they are straying outside their comfort zone and warn their human users that they need human assistance or just should not be used in such a situation. We must assign a duty to AI system designers to ensure that their creations inform users of their limitations, and specifically warn users when they are asked to operate out of their scope. AI systems must have the ability to explain their reasoning in a way that users can understand and assent to. Because of their open-ended behaviour, AI systems are also inherently hard to verify. We must develop software engineering techniques to address this. Since AI systems are increasingly self improving, we must ensure that these explanations, warnings and verifications keep pace with each AI system's evolving capabilities.

The concerns of Hawking and others were addressed in an earlier CACM Viewpoint [Dietterich & Horvitz, 2015]. While downplaying these concerns, Dietterich & Horvitz also categorise the kinds of threats that AI technology *does* pose. This apparent paradox can be resolved by observing that the various threats that they identify are caused by AI technology being too dumb, not too smart.

AI systems are, of course, by no means unique in having bugs or limited expertise. Any computer system deployed in a safety or security critical situation potentially poses a threat to health, privacy, finance, etc. That is why our field is so concerned about program correctness and the adoption of best software engineering practice. What is different about AI systems is that many people have unrealistic expectations about the scope of their expertise, simply because they exhibit *intelligence* — albeit in a narrow domain.

The current focus on the very remote threat of super-human intelligence is obscuring this very real threat from sub-human intelligence.

But could such dumb machines be sufficiently dangerous to pose a threat to humanity? Yes, if, for instance, we were stupid enough to allow a dumb machine the autonomy to unleash weapons of mass destruction. We came close to such stupidity with Ronald Reagan and Edward Teller's 1983 proposal of a Strategic Defense Initiative (SDI, aka 'Star Wars')[2]. Satellite-based sensors would detect a Soviet ballistic missile launch and super-powered x-ray lasers would zap these missiles from space before they got into orbit. Since this would need to be accomplished within seconds, no human could be in the loop. I was among many computer scientists who successfully argued that the most likely outcome was a false positive that would trigger the nuclear war it was designed to prevent. There were precedents from missile early-warning systems that had been triggered by, among other things, a moon-rise and a flock of geese. Fortunately, in these systems a human *was* in the loop to abort any unwarranted retaliation to the falsely suspected attack. A group of us from Edinburgh met UK Ministry of Defence scientists, engaged with SDI, who admitted that that they shared our analysis. The SDI was subsequently quietly dropped by morphing it into a saner programme. This is an excellent example of non-computer scientists over-estimating the abilities of dumb machines. One can only hope that, like the UK's MOD scientists, the developers of such weapon systems have learnt the institutional lesson from this fiasco. We all also need to publicise these lessons to ensure they are widely understood. Similar problems arise in other areas too, e.g., the 2010 flash crash demonstrated how vulnerable society was to the collapse of a financial system run by secret, competing and super-fast autonomous agents.

Another potential existential threat is that AI systems may automate most forms of human employment [Richard Susskind, 2015, Vardi, 2015]. If my analysis is correct then, for the foreseeable future, this automation will develop as a coalition of systems, each of which will automate only a narrowly defined task. It will be necessary for these systems to work collaboratively. Humans will be required to: orchestrate the coalition; recognise when a system is out of its depth; and deal with these 'edge cases' interactively. The productivity of human workers will be, thereby, dramatically increased and the cost of the service provided by this multi-agent approach will be dramatically reduced, perhaps leading to an increase in the services provided. Whether this will provide both job satisfaction and a living income to all humans can currently only be an open question. It is up to us to invent the future in which it will do, and to ensure that this future is maintained as the capability and scope of AI systems increases. I don't underestimate the difficulty of achieving this. The challenges are more political and social than technical, so this is a job for the whole of society.

As AI progresses, we will see even more applications that are super-intelligent in a narrow area and incredibly dumb everywhere else. The areas of successful application will get gradually wider and the areas of dumbness narrower, but not disappear. I believe this will remain true even when we do have a deep understanding of human cognition. Maggie Boden has a nice analogy with flight. We do now understand how birds fly. In principle, we could build ever more accurate simulations of a bird, but (a) this would incur an increasingly exorbitant cost and (b) we already achieve satisfactory human flight by alternative means: aeroplanes, helicopters, paragliders, etc. Similarly, we will develop a zoo of highly-diverse, AI machines, each with a level of intelligence appropriate to its task — not a new uniform race of general-purpose, super-intelligent, humanity supplanters.

[2] `https://en.wikipedia.org/wiki/Strategic_Defense_Initiative` (accessed 31^{st} December 2015).

References

[Cellan-Jones, 2014] Cellan-Jones, Rory, (December 2014). Stephen Hawking warns artificial intelligence could end mankind. BBC Interview, http://www.bbc.co.uk/news/technology-30290540.

[Davis & Marcus, 2015] Davis, Ernest and Marcus, Gary. (August 2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

[Dietterich & Horvitz, 2015] Dietterich, Thomas G. and Horvitz, Eric J. (2015). Rise of concerns about ai: Reflections and directions. *Communications of the ACM*, 58(10):38–40.

[Good, 1965] Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. *Advances in Computers*, 6.

[Kurzweil, 2005] Kurzweil, Ray. (2005). *The Singularity is Near*, pages 135–136. Penguin Group.

[Richard Susskind, 2015] Richard Susskind, Daniel Susskind. (2015). *The Future of the Professions: How Technology Will Transform the Work of Human Experts*. OUP Oxford.

[Sloman, 1995] Sloman, A. (1995). Exploring design space and niche space. In *5th Scandinavian Conf. on AI*. IOS Press, Amsterdam.

[Ulam, 1958] Ulam, Stanislaw. (May 1958). Tribute to John von Neumann. *Bulletin of the American Mathematical Society*, 64(3, part 2):1–49.

[Vardi, 2015] Vardi, Moshe Y. (2015). The future of work: but what will humans do? *Communications of the ACM*, 58(12).

[Warwick, 2004] Warwick, Kevin. (2004). *March of The Machines*. University of Illinois Press.