DOI:10.1145/2950042

Alan Bundy

Viewpoint
Smart Machines Are Not a Threat 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[1] 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[8] as the first to use the term *the singularity*[a]—the point at which artificial intelligence exceeds human intelligence. Ray Kurzweil[5] has predicted that the singularity will occur around 2045—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.'[4] The fear is that these super-intelligent machines will pose an existential threat to humanity, for example, keep humans as pets or kill us all[10]—or maybe humanity will just be a victim of evolution. (For additional information, see Dubhashi and Lappin's argument on page 39.)

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 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 modeled using a multidimensional space, with many different kinds of intelligence and with AI progressing in many different directions.[6]

AI systems occupy points in this multidimensional space that are unlike any animal species. In particular, their expertise tends to be very high in very narrow areas, but nonexistent 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 could not even move the pieces on a physical board.

▸ **Tartan Racing** was a self-driving car, built by Carnegie Mellon University and General Motors, which won

a https://en.wikipedia.org/wiki/Technological_singularity

COLLAGE BY ANDRIJ BORYS ASSOCIATES/SHUTTERSTOCK

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 could not play chess or do anything other than drive a car.

▸ **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 cannot play chess or drive a car. IBM is developing versions of Watson for a wide range of other domains, including healthcare, the pharmaceutical industry, publishing, biotechnology, and a chatterbox for toys. Each of these applications will also be narrowly focused.

▸ **AlphaGo** was a Go-playing program, developed by Google's DeepMind, that beat the World-class player, Lee Sedol, 4–1 in October 2015. 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 could not 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 *AI 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 more than 60 years ago at the 1956 inaugural AI conference at Dartmouth? It is generally agreed that one of the key enabling technologies will be commonsense reasoning. A recent *Communications* article[2] argues that, while significant progress has been made in several areas of reasoning: temporal, geometric, multi-agent, and so forth, 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 regarded as inadequate in not providing the integration of multiple cognitive processes routinely employed by humans.

Many humans tend to ascribe too much intelligence to narrowly focused AI systems.

I am not attempting to argue that AI general intelligence is, in principle, impossible. I do not believe 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 am not holding my breath. I have lived through too many AI hype cycles to expect the latest one to deliver something that previous cycles have failed to deliver. And I do not 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, for example, 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 behavior 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 recognize 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

Calendar of Events

February 4–8
PPoPP '17: 21st ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming
Austin, TX,
Sponsored: ACM/SIG,
Contact: Vivek Sarkar,
Email: vsarkar@rice.edu

February 6–10
WSDM 2017: 10th ACM International Conference on Web Search and Data Mining
Cambridge, U.K.,
Co-Sponsored: ACM/SIG,
Contact: Milad Shokouhi,
Email: milads@microsoft.com

February 21–22
HotMobile '17: The 18th International Workshop on Mobile Computing Systems and Applications
Sonoma, CA,
Sponsored: ACM/SIG,
Contact: Elizabeth M. Belding,
Email: ebelding@cs.ucsb.edu

February 22–24
FPGA '17: The 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays
Monterey, CA,
Sponsored: ACM/SIG,
Contact: Jonathan Greene,
Email: jonathan.greene@microsemi.com

February 25–March 1
CSCW '17: Computer Supported Cooperative Work and Social Computing
Portland, OR,
Sponsored: ACM/SIG,
Contact: Steven E. Poltrock,
Email: spoltrock@gmail.com

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 behavior, 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 these explanations, warnings, and verifications keep pace with each AI system's evolving capabilities.

The concerns of Hawkings and others were addressed in an earlier *Communications* Viewpoint by Dietterich and Horvitz.[3] While downplaying these concerns, Dietterich and Horvitz also categorize the kinds of threats that AI technology *does* pose. This apparent paradox can be resolved by observing that the various threats 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, and other realms. 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 some people may 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').[b] 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 moonrise 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 U.K. Ministry of Defence scientists, engaged with SDI, who admitted they shared our analysis. The SDI was subsequently quietly dropped by morphing it into a saner program. This is an excellent example of non-computer scientists overestimating the abilities of dumb machines. One can only hope that, like the U.K.'s MOD scientists, the developers of such weapon systems have learned the institutional lesson from this fiasco. We all also need to publicize these lessons to ensure they are widely understood. Similar problems arise in other areas too, for example, 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.[7,9] 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, with humans: orchestrating the coalition, recognizing when a system is out of its depth and dealing with these 'edge cases' interactively. The productively 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 this future is maintained as the capability and scope of AI systems increases. I do not 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 this would incur an increasingly exorbitant cost and we already achieve satisfactory human flight by alternative means: airplanes, helicopters, paragliders, and so forth. 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. C

b https://en.wikipedia.org/wiki/Strategic_Defense_Initiative

References

1. Cellan-Jones, R. Stephen Hawking warns artificial intelligence could end mankind. BBC Interview, (Dec. 2014); http://www.bbc.co.uk/news/technology-30290540.
2. Davis, E. and Marcus, G. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM 58*, 9 (Sept. 2015), 92–103.
3. Dietterich, T.G. and Horvitz, E.J. Rise of concerns about AI: Reflections and directions. *Commun. ACM 58*, 10 (Oct. 2015), 38–40.
4. Good, I.J. Speculations concerning the first ultra-intelligent machine. *Advances in Computers 6* (1965).
5. Kurzweil, R. *The Singularity is Near.* Penguin Group, 2005, 135–136.
6. Sloman, A. Exploring design space and niche space. In *Proceeding of the 5th Scandinavian Conference on AI*. IOS Press, Amsterdam, 1995.
7. Susskind, R. and Susskind, D. *The Future of the Professions: How Technology Will Transform the Work of Human Experts.* OUP Oxford, 2015.
8. Ulam, S. Tribute to John von Neumann. *Bulletin of the American Mathematical Society 64*, 3, part 2) (May 1958), 1–49.
9. Vardi, M.Y. The future of work: but what will humans do? *Commun. ACM 58*, 12 (Dec. 2015).
10. Warwick, K. *March of The Machines.* University of Illinois Press, 2004.

Alan Bundy (A.Bundy@ed.ac.uk) is Professor of Automated Reasoning at the School of Informatics, University of Edinburgh, Scotland.

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