Our Skynet Moment: Debating Morality of AI

Jayakumar, Shashi

2016

Jayakumar, S. (2016). Our Skynet Moment: Debating Morality of AI. (RSIS Commentaries, No. 137). RSIS Commentaries. Singapore: Nanyang Technological University.

https://hdl.handle.net/10356/80991

RSiS
Nanyang Technological University
S. RAJARATNAM SCHOOL OF INTERNATIONAL STUDIES

RSIS COMMENTARY

www.rsis.edu.sg

No. 137 – 7 June 2016

Our Skynet Moment: Debating Morality of AI

By Shashi Jayakumar

Synopsis

The rapid growth of artificial intelligence (AI) has serious implications for our future. The issues and their oversight are not just the domain of computer engineers, technologists and AI experts. Policymakers, Smart Nation experts and security officials too should come together with them to ponder implications and set out the parameters, if needed, for future research and development.

Commentary

IN MARCH this year, AlphaGo, a machine created by Google's artificial intelligence (AI) arm, DeepMind, trounced Lee Sedol, a grandmaster at Go, the ancient Chinese game. AlphaGo used cutting-edge AI to beat a player acknowledged to be one of the greatest ever.

AlphaGo’s achievement comes almost 20 years after IBM’s Deep Blue beat reigning world chess champion Gary Kasparov. Deep Blue used brute-force calculation and sheer computing power. AlphaGo is a far more complex machine utilising deep neural networks and reinforcement learning, independent of human input. The machine learnt on its own as it progressed and got stronger as it played.

Significance of AlphaGo

What AlphaGo has shown is that advances in AI once thought to need several decades to be made can be compressed into a few years. Change is happening at a very fast rate and policymakers may not have the luxury of time to adjust and to make decisions. It is time to start thinking about what exactly this all means for us as individuals and for humanity as a whole.

Certainly, AI will benefit us all in ways that we are only just beginning to fathom. Consider, for example, AI as a tremendous force for good in national security. Machine learning tools have already been applied to complex security situations around the world. There have, for example, been thought-provoking trials on modeling the behaviour of the Islamic State in Iraq and Syria, telling us (through the application of AI to big data) far better than most analysts could where the militant group might plant improvised explosive devices.

The Singapore security architecture already has systems that parse big data and weak signals, such as the Risk Assessment and Horizon Scanning System. Could its predictive capacities be improved with the use of AI? This would not, of course, be a silver bullet to predict when a terrorist attack might occur, but a system that learns what constitutes good predictions could in theory become increasingly proficient at avoiding bad ones, across a whole variety of scenarios. It will not find us the proverbial needle, but at least analysts might have more definition in terms of which haystack to look in.

AI would also be of immeasurable benefit in the new initiative announced by the Home Team - SG Secure - which would have an element of surveillance and the use of closed-circuit television (CCTV) cameras. The smartest CCTV systems are already beginning to use some elements of AI. These obviate the need for tedious and cumbersome cross-checking of profiles and reams of data, and, in some trials, have identified patterns that point to a likely crime even before the act itself.

When AI Gets a Mind of Its Own

Strong AI systems are capable of learning autonomously, improving their capabilities with each iteration. This has tremendous implication, as it suggests that a system initially set with a defined utility or purpose might "learn" to develop a utility or purpose different from what the designers intended.

AI researchers have reached the point of being prepared to seriously discuss whether this recursive drive for improvement and resource acquisition inherent in strong AI systems may ultimately mean that a machine's real concept of utility might diverge at some point from what was intended.

The machine might still be designed by a human agent, but it might not be designed well enough. AI systems may seek to obtain more resources for whatever goals they might have. And in doing so, it is possible that an AI system would develop new processes to complete tasks faster and become more capable than it was designed to be. An AI system could essentially "re-architect" itself.

In popular culture, the reference point would be the moment in an early 1990s Hollywood movie, etched into the consciousness of the generation that grew up with it, when Skynet, an AI system created by humans, gains self-awareness and resists human attempts to disable it. The result is a world war that leads to humankind's near annihilation, with humans barely clinging on in a dystopian future world ruled by killer robots.

Skynet Moment

AI experts are right, of course, to say we are nowhere near a Skynet moment. Notwithstanding this, some experts have argued that that there needs to be a set of well thought-out countervailing instructions (conceivably a type of deep-lying circuit breaker or "kill" switch) embedded in all AI systems, simply because a self-improving, strong AI system would go to lengths that we cannot fully comprehend in order to fulfill its goals.

And AI systems might in some circumstances be able (say, because of poorly designed controls) to break their constraints. Therefore, so the argument runs, self-improvement capacity (or the drive to gain more resources) might have to be limited in a coherent and well-thought-out manner, where it is impossible to circumvent those constraints.

The debate is not simply a theoretical one confined to arcane journals. It affects, for example, ongoing debates on lethal autonomous drones. Alone among the major powers, the United States requires human input before lethal force is exercised by its drones. But for how much longer?

As machine learning algorithms improve, we might reach a situation where retaining the human element means a loss of efficiency - removing the human decision-making might give a system the edge over its targets (which might be humans or other drones). But how would we know that the system is making the right decisions?

Debate on Morality for AI and Robots

There is, in fact, an ongoing debate about the issue of building morality into AI systems and robots. Research on how to build a sense of right and wrong and moral consequence into autonomous robotic systems, including lethal autonomous weapon systems, is already being funded.

Whether this can ever be successfully done remains to be seen. Code instructions are deterministic in nature, whereas the understanding of inappropriate or wrong behaviour is subjective.

In the absence of certainty on these issues, many experts, chief executive officers and futurists are calling for commonly recognised norms to regulate advanced AI research, or even for a pause in such research until a coherent framework can be developed.

Within Singapore, technologists, academics, social scientists and the Government should come together in communities of practice to discuss approaches to these issues. This is already beginning to happen. It would be useful for all those concerned to be mindful of widely-accepted principles that enshrine, within all experimentation, the primary responsibility to humankind, with this responsibility privileged even above the interests of science and research.

Shashi Jayakumar is Senior Fellow and Head, Centre of Excellence for National Security, S. Rajaratnam School of International Studies (RSIS), Nanyang Technological University, Singapore. An earlier version appeared in The Straits Times.

Nanyang Technological University
Block S4, Level B4, 50 Nanyang Avenue, Singapore 639798
Tel: +65 6790 6982 | Fax: +65 6794 0617 | www.rsis.edu.sg