AI Risk Mitigation through Democratic Governance

Colin K Garvey
PhD Candidate, Science & Technology Studies Department, Rensselaer Polytechnic Institute
Sage Labs 5710, 110 8th Street, Troy, NY, 12180 USA
garvec@rpi.edu

Abstract

This dissertation project asks two fundamental questions: What are the risks of AI? And what can be done about them? My research goes beyond existential threats to humanity to consider seven dimensions of AI risk: military, political, economic, social, environmental, psychophysiological, and spiritual. I examine extant AI risk mitigation strategies and, finding them insufficient, use a democratic governance framework to propose alternatives. In this paper, I outline the project and discuss four risk dimensions.

Introduction

Concern about the negative social impacts of AI has been growing in recent years as rapid technological developments bring the promises and threats of AI closer to reality. On the one hand, these concerns have been stoked by some high-profile figures in the tech industry, such as Elon Musk, who tweeted that AI is "more dangerous than North Korean nukes." On the other, Mark Zuckerberg and others have defended AI as essentially risk-free. In the frenzied media coverage of this debate, more heat than light has been generated as most people are still left wondering if AI is dangerous, or not. This uncertainty highlights the importance of the following two questions: What are the risks of AI? And what should be done about them? These are the two questions my dissertation research seeks to answer.

Overview of the Dissertation Project

The common framing of AI impacts in terms of ambivalent extremes—either utopia or dystopia, heaven or hell—impairs our ability to understand the risks of this emerging technology. Utopians see no risk, while dystopians see only *existential* risk (Bostrom 2014; Müller 2016), the danger that AI will somehow make humanity extinct. This absolutism leaves little room to steer AI toward robustly beneficial futures for a majority of humanity: either nothing needs to be done, or nothing can be done.

My dissertation project attempts to disrupt this dichotomous framing by articulating seven dimensions of AI risk, four of which are described in the next section. However, by taking a social scientific approach to risk, I focus on the *who* as much as the *what*. Through participant observation at AI conferences and semi-structured interviews with experts, I seek to understand how AI scientists, developers, entrepreneurs, funders, and users are creating risks, what those risks are, and who is being put at risk by AI.

The second part of the project examines strategies for risk mitigation I have observed in my fieldwork. How well does the present system of governance cope with risk? By focusing on narrowly technical questions of *safety*, these strategies largely ignore the economic, social, and political contexts in which the decision making processes leading to risky AI take place. Because they fail to address broader issues of governance, I argue that purely technical approaches to risk mitigation are insufficient measures.

Therefore, the third part of my dissertation project explores how the political structure of the decision making processes in AI research and development (R&D) contribute to risk. What changes to governance structures might help mitigate the scope and magnitude of such risks? Here I employ a framework described in my AAAI Student Abstract to identify barriers to more intelligent, democratic governance of AI, and propose strategies for overcoming these barriers. Finally, I consider how the case of AI risk governance can inform the framework itself.

Methods

Data sources analyzed for this project include: primary documents from AI-focused institutions and tech companies; AI policy documents from governments and private organizations; interviews with technical experts, social scientists, and laypeople; as well as participant observation at AI conferences and laboratories in the USA and Japan.

The State of Risk in AI

What types of AI risk are being created? Historically the field of AI paid little attention to risk (Barrat 2013). This changed in 2014 when Stephen Hawking, in a series of op-eds, began sounding the alarm about the existential risk to humankind posed by AI. Others quickly followed suit (Anderson 2014), and these initial conditions locked the emerging conversation into a trajectory that stifled more nuanced views even as the issue gained media attention.

Seven Dimensions of AI Risk

This project seeks to disrupt this singular focus on existential risk by expanding the categories under consideration to include the 1) *military*, 2) *political*, 3) *economic,* 4) *social,* 5) *environmental*, 6) *psychophysiological*, and 7) *spiritual* risks of AI. Here I describe the first four dimensions.

Military Risks

Bracketing *Terminator*-like scenarios altogether, the military applications of AI still pose serious risks to humanity Led by the USA, China, and Russia, national militaries are producing a new generation of Autonomous Weapons Systems (AWSs). Proponents argue they will save lives, but these new systems are likely to introduce many new problems. Journalists are already using the term "arms race" to describe China's massive investments in AI (Kania 2017).

Political Risks

AI technologies provide unprecedented tools for elites to manipulate opinion and exploit have-nots (O'Neil 2016). The 2016 US presidential election provided a powerful example. The AI behind Facebook's newsfeeds and Google's search results led to partisan isolation, keeping voters in private "echo chambers"; right-wing groups used AI to rapidly disseminate "fake news" and divisive messages designed to stoke suspicion of certain ethnic and religious groups; new modeling techniques allowed for "micro-targeting" of the specific demographics most susceptible to manipulation. AI risks powering the "post-truth era" that has thrown America's democracy into crisis.

Economic Risks

Many have argued AI threatens jobs (Brynjolfsson and McAfee 2012; Ford 2015; Kaplan 2015). The most-cited figure is that "47% of the US workforce is at risk of automation" (Frey and Osborne 2013). Although this has been criticized and other studies reduced the number to 9% across OECD countries (Arntz, Gregory, and Zierahn 2016), the wide variation in quantitative estimates highlights experts' uncertainty about the economic risks of AI.

Social Risks

Because most modern AI relies on human-generated data for learning, it systematically reproduces biases in that data (Caliskan, Bryson, and Narayanan 2017). Therefore, AI risks automating and entrenching discriminatory social practices. Algorithmic discrimination has already been reported in criminal sentencing (Angwin et al. 2016) and a wide variety of other contexts (Crawford 2016).

Significance and Impact

My hope is that by providing a more nuanced categorization of the risks of AI and expanding the range of mitigation strategies that could be harnessed to cope with them, this project will draw more stakeholders into discussions about AI risk, open possibilities for new modes of governance, and improve outcomes by facilitating risk mitigation.

Conclusion

Mitigating some or even all of the risks described in my dissertation may require significant changes to the decision making processes currently governing AI R&D. Yet by better aligning those processes with the social values of modern democracies, such changes may not only reduce risk, but help to ensure that AI benefits society as well.

References

Anderson, Lessley. Oct 2014. "Elon Musk's Artificial Intelligence Fear: Machines Could Wipe Us Out." *Vanity Fair*.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." *ProPublica*.

Barrat, James. 2013. *Our Final Invention*. New York: Thomas Dunne Books.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Brynjolfsson, Erik, and Andrew McAfee. 2012. *Race against the Machine*. Lexington, Mass.: Digital Frontier Press.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science* 356 (6334):183–186.

Crawford, Kate. June 25, 2016. "Artificial Intelligence's White Guy Problem." *The New York Times*.

Ford, Martin. 2015. *Rise of the Robots: Technology and the Threat of a Jobless Future*. New York: Basic Books, a member of the Pereus Books Group.

Frey, Carl Benedikt, and Michael A. Osborne. 2013. "The Future of Employment: How Susceptible Are Jobs to Computerisation." Oxford Martin School of Business.

Kania, Elsa. March 9, 2017. "The Next U.S.-China Arms Race: Artificial Intelligence?" Text. *The National Interest*.

Kaplan, Jerry. 2015. *Humans Need Not Apply*. New Haven: Yale University Press.

Müller, Vincent C., ed. 2016. *Risks of Artificial Intelligence*. Boca Raton, FL: CRC Press.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books.