AI Risk Mitigation Through Democratic Governance

Introducing the 7-Dimensional AI Risk Horizon

Colin Garvey
Department of Science & Technology Studies
Rensselaer Polytechnic Institute
Troy, New York, USA
garvec@rpi.edu

ABSTRACT

My dissertation asks two fundamental questions: What are the risks of AI? And what should 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. This paper outlines the project and introduces the risk dimensions.

ACM Reference format:

Colin Garvey. 2018. AI Risk Mitigation Through Democratic Governance: Introducing the 7-Dimensional AI Risk Horizon. In *AIES '18: AAAI/ACM Conference on AI, Ethics, and Society Proceedings (AIES'18), February 2-3, 2018, New Orleans, LA, USA,* ACM, NY, NY, USA, 3 pages
https://doi.org/10.1145/3278721.3278801

1 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. Some high-profile figures in the tech industry have spoken out to validate these fears, while others have defended AI as essentially risk-free. The truth is surely somewhere in between, yet more heat than light has been generated in this debate, leaving the public to wonder: Is AI dangerous, or not? This uncertainty raises at least two questions. What *are* the risks of AI? And what should be done about them?

2 OVERVIEW OF THE DISSERTATION

The common framing of AI impacts in terms of ambivalent extremes—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," the danger that AI will somehow make humanity extinct [3]. This absolutism leaves little room to steer AI toward robustly beneficial futures: either nothing needs to be done, or nothing can be done.

AIES '18, February 2–3, 2018, New Orleans, LA, USA

ACM ISBN 978-1-4503-6012-8/18/02.
https://doi.org/10.1145/3278721.3278801

My project disrupts this dichotomous framing by articulating seven dimensions of AI risk. In each, I focus on the *who* as much as the *what*. Who is put at risk? Which socioeconomic groups are at greatest risk? Least risk? Who is creating these risks?

Part 1 answers these questions by using the 7-dimensional risk horizon to guide an empirical examination of *who* is being put at risk, *what* the risks are, and *how* AI scientists, developers, entrepreneurs, funders, and users are creating those risks.

Part 2 then asks, How is risk currently governed in AI? The risk mitigation strategies observed in my fieldwork ignore the economic, social, and political contexts of the decision making processes leading to risky AI. Because they fail to address macro-level contextual issues, these purely technical approaches to risk mitigation are insufficient. My *socio*technical approach reveals how risks emerge from the non-democratic political structure of the decision making processes in AI research and development.

Part 3 thus asks, What changes to the governance of AI R&D might help mitigate these risks? Here I employ a democratization framework described elsewhere [10] to identify barriers to more democratic governance of AI and propose strategies for overcoming those barriers. Finally, I consider how the specific case of AI risk governance can inform the framework itself.

2.1 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 AI and robotics laboratories in the USA and Japan.

2.2 Introducing the 7-Dimensional Risk Horizon

Historically, the field of AI has paid little attention to risk [2]. However, recent years have seen the "existential risk" of AI receive considerable media coverage and scholarly attention. Unfortunately, this development narrowed the focus onto extreme scenarios rooted in science fiction and locked the emerging discussion of AI risk into a dichotomous trajectory that stifled more nuanced views even as the topic itself grew in popularity.

My dissertation disrupts this singular focus on existential risk by articulating seven pragmatic dimensions of AI risk: *military*, *political*, *economic*, *social*, *environmental*, *psycho-physiological*, and *spiritual*. These seven dimensions, sketched below,

constitute a "risk horizon" that should prove useful for AI futures scanning.

2.2.1 Military Risks. Ignoring sci-fi scenarios (e.g. *Terminator*) entirely, the military applications of AI still pose serious risks to humanity. Led by the USA and China, national militaries are producing a new generation of Autonomous Weapons Systems (AWS). Proponents claim they will save lives, but AWSs will introduce new problems as well, such as military arms races [1].

2.2.2 Political Risks. AI and Big Data provide unprecedented tools for elites to manipulate popular opinion and exploit have-nots [15]. This was shown dramatically in the 2016 US presidential election. The AI behind newsfeeds and search results led to partisan isolation, locking voters into party-affiliated echo chambers. Moreover, right-wing groups disseminated "fake news" and divisive messages by using AI to achieve unprecedented levels of fine-grained "micro-targeting" of specific demographics. AI thus risks accelerating and empowering the "post-truth era" that has thrown democracy in the US and elsewhere into crisis.

2.2.3 Economic Risks. Many experts agree that AI threatens jobs [8, 14]. However, they disagree over just how severe the threat is. The most-cited figure is that "47% of the US workforce is at risk of automation" [9]. Subsequent studies challenge this estimate, but the wide variation in predictions highlights experts' uncertainty about the size and scope of AI's economic impact. The uncomfortable truth is that no one knows what will happen.

2.2.4 Social Risks. AI trained on human-generated data systematically reproduces biases in that data [5]. AI thus risks entrenching discriminatory social practices that disproportionately impact have-nots. Algorithmic harms have already been identified in multiple social contexts [6]. Scientific practice could also be impacted. Lacking causal models, the use of machine learning in medicine could lead to an era of "digital phrenology" [11].

2.2.5 Environmental Risks. The physical environment is typically overlooked in analyses and predictions about the impact of technology on human life [7]. As AI is adopted in cars, homes, and the workplace, it risks further alienating people from their environment and each other [4]. Moreover, AI, like BitCoin, constitutes an infinite sink for energy. As industry and consumer use of AI grows, will the pace of resource extraction and destruction of the natural environment increase, or decelerate?

2.2.6 Psycho-physiological Risks. In the "attention economy," tech companies succeed by producing addictive products with negative effects on user health. The mere presence of one's smartphone can reduce cognitive capacity [17]. The current epidemic of teen depression and suicide correlates with screen-time and social media use [16]. Yet many propose more computer education from an earlier age as the only means of coping with AI-induced job loss in the attention economy [8]. AI thus risks disempowering not only workers, but their children as well.

2.2.7 Spiritual Risks. AI raises questions about the place and shape of human nature in an increasingly automated world [12]. Spiritual traditions across time and around the world uphold meditative states of hypostatic awareness as key to accessing the transcendent, eternal aspects of ourselves [13]. As we co-evolve with intelligent machines, will our capacity to reflect on the mysteries of Being and the Beyond be enhanced or diminished?

3 SIGNFICANCE AND IMPACT

The 7-dimensional risk horizon can scaffold more nuanced understandings AI risk and expand the range of mitigation strategies that could be harnessed to cope with them. Broader risk awareness might draw more stakeholders into discussions about AI, open possibilities for new modes of governance, and improve overall outcomes by facilitating risk mitigation at multiple scales.

4 CONCLUSION

Mitigating even some 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 democratic societies as well.

REFERENCES

[1] Julian E. Barnes and Josh Chin. March 2, 2018. "The New Arms Race in AI." *Wall Street Journal*, sec. Life.

[2] James Barrat. 2013. *Our Final Invention: Artificial Intelligence and the End of the Human Era.* Thomas Dunne Books, New York.

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

[4] Rodney Brooks. 2017. Robotic Cars Won't Understand Us, and We Won't Cut Them Much Slack. *IEEE Spectrum* 54(8): 34–51.

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

[6] Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. 2017. AI Now 2017 Report. AI Now Institute, New York University.

[7] Nathan Ensmenger and Rebecca Slayton. 2017. Computing and the Environment: Introducing a Special Issue of *Information & Culture*. *Information & Culture* 52(3): 295–303.

[8] Martin Ford. 2015. *Rise of the Robots.* Basic Books, New York.

[9] Carl Benedikt Frey and Michael A. Osborne. 2013. The Future of Employment. Oxford Martin School of Business, University of Oxford.

[10] Colin Garvey. 2018. A Framework for Evaluating Barriers to the Democratization of Artificial Intelligence. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18).* AAAI, New Orleans, LA, USA, 8079-8080.

[11] ---. 2018. Interview with Colin Garvey, Rensselaer Polytechnic Institute. Artificial Intelligence and Systems Medicine Convergence. *OMICS: A Journal of Integrative Biology* 22(2): 130–32.

[12] Robert M Geraci. 2010. *Apocalyptic AI: Visions of Heaven in Robotics, AI, and Virtual Reality.* Oxford University Press, New York.

[13] William S. Haney. 2006. *Cyberculture, Cyborgs and Science Fiction: Consciousness and the Posthuman.* Rodopi, New York.

[14] Jerry Kaplan. 2015. *Humans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence.* Yale University Press, New Haven.

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

[16] Jean M. Twenge, Thomas E. Joiner, Megan L. Rogers, and Gabrielle N. Martin. 2018. Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time. *Clinical Psychological Science* 6(1): 3–17.

[17] Adrian F. Ward, Kristen Duke, Ayelet Gneezy, and Maarten W. Bos. 2017. Brain Drain: The Mere Presence of One's Own Smartphone Reduces Available Cognitive Capacity. *Journal of the Association for Consumer Research* 2(2): 140–54. https://doi.org/10.1086/691462.