Control and Responsible Innovation in the Development of AI and Robotics

Executive Summary

Edited by Wendell Wallach

With contributions from Carla Gomes, Gary Marchant, David Roscoe, Francesca Rossi, Stuart Russell, Bart Selman, and Shannon Vallor

A project of The Hastings Center, supported by funding from the Future of Life Institute

Executive Summary/Recommendations

Artificial intelligence (AI) and robotics promise tremendous benefit but are accompanied by an array of legal, safety, and ethical concerns. These fields of research are already beginning to transform every sector of modern life. According to a 2017 report from PricewaterhouseCoopers, AI-driven GDP growth worldwide will be $15.7 trillion by 2030. As we reap the benefits and productivity gains of AI and robotics, how can we minimize risks and undesirable societal impacts through engineering, ethics, and oversight?

Recognizing that the challenges posed by AI and robotics are too great to be fully addressed by experts in any one field, The Hastings Center's project "Control and Responsible Innovation in the Development of AI and Robotics," funded by the Future of Life Institute, convened a series of three workshops bringing together leading AI developers and experts from many other fields working on issues and approaches related to the ethics and governance of autonomous systems. In these transdisciplinary, two-day workshops, 20 to 25 participants convened to share and receive feedback on the issues or technical challenges they faced in their work.

Research, innovation, and the deployment of AI and robotic systems are proceeding rapidly, and so, too, is the emergence of a transdisciplinary community of researchers in AI and the social sciences dedicated to AI safety and ethics. The Hastings AI workshops played a seminal role in catalyzing the emergence of this worldwide network of organizations and individuals. Many of the workshop participants are recognized as leaders in that network. The Hastings Center workshops role in catalyzing the emergence and evolution of an AI safety community is the major output of the project.

Four themes were central to the project's work:

A) Silo-busting: How might transdisciplinary collaboration to solve the various challenges be facilitated? While the experts attending these workshops knew those in their own field, they initially did not know those from other fields working on similar problems, nor did they understand the research in these potentially complementary fields.

B) Value alignment and machine ethics: Can computational systems that honor human values in their choices and actions be designed and engineered? Aligning the values and goals of advanced forms of AI with those of humans was proposed by leaders within the AI research community as a means to ensure the safety and beneficial results of future superintelligent machines. A field directed at implementing sensitivity to ethical considerations in bots and robots and factoring those into the system's choices and action had already been progressing for more than a decade. The latter field is commonly referred to as machine ethics, machine morality, or the development of moral machines. Key participants in machine ethics approaches and value alignment approaches were brought together at The Hastings workshops to explore possible ways to collaborate together.

C) Shorter-term safety concerns versus long-term challenges: Does work on nearer-term challenges lay foundations for ensuring the safety or controllability of AGI or are the challenges

posed by advanced systems of a totally different order? While some participants were focused on ensuring the safety or controllability of future artificial general intelligence or superintelligence (AGI or ASI), many other participants directed their work at the safety of systems designed to address more near-term tasks.

D) Comprehensive and agile governance: During the development of AI, what forms of ethical or legal oversight should be put in place to monitor progress, flag gaps, coordinate the activities of the many organizations and governmental bodies jumping into this space, and facilitate multistakeholder dialogue? Discussions in workshops I and III elaborated upon a model of governance that embraces hard law and regulations, soft governance (standards, laboratory practices, insurance policies, etc.), industry self-governance, and technological solutions such as machine ethics and value alignment.

Drawing on the discussions over the course of the project, three core recommendations emerged.

1) A consortium of industry leaders, international governmental bodies and nongovernmental institutions, national and regional (e.g., the European Union) governments, and AI research laboratories should convene an International Congress for the Governance of AI (ICGAI) by November 2019. This Congress will initiate the creation of a new international mechanism for the agile and comprehensive monitoring of AI development and any gaps in oversight that need to be addressed. In determining appropriate methods for addressing gaps it will consider technical solutions, procedures for responsible innovation by corporations and research laboratories, and standards and soft law. Given difficulties in enacting hard law and regulatory solutions and of changing laws as circumstances change, hard law and regulations will be appropriate only when other solutions are insufficient. Certainly, some laws and regulations must be enacted to deter dangerous practices, protect rights, and to enforce egregious violations of established standards. A first meeting to plan for this proposed International Congress was convened in September 2018 in New York City when the UN General Assembly was in session.

2) Universities and colleges should incentivize the education of a cadre of polymaths and transdisciplinary scholars with expertise in AI and robotics, social science research, and philosophy and practical ethics. Foundations and governmental sources of funding should contribution to the establishment of transdisciplinary research centers. In particular, foundations and governments should fund centers dedicated to forging methods to implement sensitivity to human values in computer systems. Various research groups have proposed a broad array of approaches to what is called the “value alignment” problem and the creation of moral machines. It is essential to fund as many of these approaches as possible in the hope that effective solutions will emerge and develop.

3) Foundations and governmental sources of funds should help establish in-depth and comprehensive analyses of the benefits and issues arising as AI is introduced into individual sectors of the economy. We identified AI and health care as a good starting point. The benefits of AI for health care are commonly touted, but what will be the tradeoffs as we implement various approaches to reaping those benefits? This deep-dive would encompass AI and health care systems, pharmaceutical and health care research, clinical practice, and public health.