Control and Responsible Innovation in the Development of AI and Robotics

Draft Final Report

Edited by Wendell Wallach

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

Gregory Kaebnick and Isabel Bolo provided editorial assistance with this report.

future of life INSTITUTE

The Hastings Center

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

Table of Contents

Section I. 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.

Section II. Introduction

Over the past five years, breakthroughs in machine learning (ML) approaches to the development of artificial intelligence (AI) have been accompanied by great excitement over AI's benefits, but also concern about serious risks and undesirable societal consequences. Beginning in the fall of 2015, **The Hastings Center** hosted a series of three transdisciplinary experts' workshops directed at ameliorating risks while maximizing progress toward reaping rewards from research in AI and robotics. The workshops were funded by a grant from the **Future of Life Institute**. The participants were leaders in many fields of research and included scientists, engineers, social and legal theorists, cognitive scientists, disaster analysts, practical ethicists, and philosophers. During their discussions, four key themes emerged:

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? Workshop I and III discussed and 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.

In addition to these four themes, many other issues were addressed, including guidelines for the deployment of AI systems that effect children, challenges to human identity, technological unemployment, lethal autonomous weapons (LAWS), AI and health care, problems arising in the coordination of humans and computational systems, and the management of complex adaptive systems. This report will cover findings and proposals emerging from the workshop discussions on the four core themes, and touch upon a few other subjects. Theme B) Value alignment and machines, and theme C) Shorter-term safety concerns versus Long-term challenges will be reviewed together in Section IV of this report.

Background and Project Origin

Breakthroughs in machine learning approaches to artificial intelligence (AI), particularly the approach known as deep learning, generated enthusiasm for untold benefits but also rekindled concern as to whether future advanced forms of AI would be safe and controllable. As public figures such as Stephen Hawking, Elon Musk, and Bill Gates weighed in on potential risks that artificial general intelligence (AGI) might pose, the Future of Life Institute (FLI) convened a January 2015 meeting in Puerto Rico of leading figures in the field of AI to address the challenges. In addition to AI researchers, a few social scientists, legal theorists, and philosophers were invited. The safety of AGI was the primary agenda for the Puerto Rico gathering, but the meeting's organizers were also interested in potential loss of jobs due to increasing automation afforded by smart systems. In addition, the program included a presentation about meetings of the Convention on Certain Conventional Weapons at the UN in Geneva on a proposed ban of lethal autonomous weapons, sometimes referred to as "killer robots."

While the AI researchers saw the Puerto Rico gathering as one of the first to include those from other fields, Wendell Wallach, who was in attendance, was struck by the paucity of participants working in the fields of machine ethics, engineering ethics, resilience engineering, joint cognitive systems, and risks posed more generally in managing complex adaptive systems, including "normal" or "system accidents." The emphasis in these fields, however, is more toward nearer-term safety and ethical challenges than toward the safety and control of AGI. Evidently the organizers knew little about the work in these other fields on AI and robotic safety, or did not consider that work a high priority.

During the Puerto Rico meeting, Elon Musk committed ten million dollars to FLI to fund research on AI safety, and that commitment was later supplemented by other funding. In concert with Stuart Russell (UC Berkeley) and Bart Selman (Cornell), Wendell Wallach (Yale) submitted a proposal to FLI to fund a series of workshops that would bring together those in the many fields that work on various facets of machine safety. Gary Marchant (ASU) was added to the project as a fourth co-chair to round out expertise in the governance of emerging technologies. The proposal was submitted to FLI by the bioethics think tank The Hastings Center, for which Wallach is a Senior Advisor.

The Hastings Center proposal started with the recognition that a vast array of challenges entailed in designing, engineering, and implementing demonstrably beneficial, safe, controllable, robust, and resilient AI systems were being addressed by scholars working on distinct research trajectories across many disciplines. Given that these engineers and researchers were often unaware of efforts in complementary fields, they lost opportunities for creative synergies, missed gaps in their own research, and occasionally reproduced the work of potential colleagues.

Bringing leaders in the various fields together to learn from each other was the first goal of the project. In addition, these three solution-directed workshops would: 1) address transdisciplinary questions, 2) develop collaborative strategies and research projects, and 3) outline first steps in a comprehensive plan that ensured autonomous systems will be demonstrably beneficial and that this innovative research progresses in a responsible manner.

The Future of Life Institute awarded The Hastings Center with a grant to fund the series of experts' workshops. The researchers and scholars who participated in these transdisciplinary sessions are listed in the Appendix.

The Hastings Center has fifty years of experience in convening workshops that bring together individuals from diverse backgrounds and differing fields of understanding, and who often identify with conflicting belief systems, to respectfully work together to solve concrete challenges. Many of the challenges addressed over the years have been the result of innovative technologies. The three expert workshops were designed so as to facilitate opportunities for the attending experts to communicate across silos and boundaries. In knitting the various fields of research closer together, we also believe there will be a carryover from our meetings to a cross-fertilization of ideas among the various colleagues of the participating experts. In effect, we hope to catalyze a process that would transform initially tenuous connections between the respective fields into a highway of collaboration.

This report captures some of the findings and recommendations that emerged from these expert workshops. In a larger sense, however, it is difficult to capture the breadth of our discussions and the impact of such seminal gatherings. Since the Puerto Rico FLI meeting and the first Hastings AI workshop, much has happened. AI is a hot topic, and there are now a broad array of conferences, workshops, and policy conclaves on the subject, many of which include participants from many fields of research. Furthermore, understanding of many of the issues has evolved, just as it did during Hastings' workshops. The role that our workshops played in that evolution is impossible to measure.

As the AI safety and ethics ecosystems has emerged, grown, and evolved, so too have the roles and responsibilities of many of the workshop participants. Examples of the increasing prominence in this ecosystem of a few workshop participants will be mentioned in the concluding section of this report. We believe the Hastings workshops played an important role in the depth and breadth of understanding of these leaders.

In other words, the Hastings workshops were an integral part of the emerging and evolving culture focused on safety, societal impacts, and the responsible development of AI. Arguably, they played a catalytic and seminal role in the early stages of cross-disciplinary work on ensuring safe and ethical AI.

During the evolution of the AI safety culture, an array of new issues, concerns, research trajectories, and policy initiatives emerged. For example, algorithmic bias and the transparency of algorithms were two subjects that received little or no attention at the FLI Puerto Rico meeting and the first Hastings AI workshop. Over the next two years, however, these two subjects took on central importance for ensuring the safety and reliability of machine learning systems. Next, the manipulation of voting behavior through social media, and facilitated by AI, moved to the fore as an issue of public concern. This in turn focused more attention on cyberwarfare and whether an international treaty limiting its expansion could be passed. More recently, adversarial techniques for undermining the reliability of learning algorithms is receiving serious attention from experts in the field. All of these issues have given rise to forging technical tools to address more short-term ethical and legal concerns and to new proposals for standards, best practices, and governmental regulation. In some cases, governmental policies are being implemented, most significantly the EU's General Data Policy Recommendations (GDPR) implemented in 2018.

Other groups that have taken on importance in the AI safety and ethics culture, such as the Partnership on AI and AI Now, are releasing extensive reports on standards, algorithmic bias, transparency, and defusing the use of AI for behavioral manipulation through social media. We of course entertained early discussions of those topics, but we will not repeat in this report subjects that are now handled in great depth elsewhere. So as to focus our reflections together, a few topics were tabled. The tabled topics included technological unemployment, lethal autonomous weapons, and explicitly legal concerns. Many of the workshop participants had expertise in these subjects that were tabled, and these topics were occasionally broached in our discussions. But we mutually agreed that these topics were being given more serious attention by other groups and would be distracting from subject on which we could make some headway.

The remainder of this report covers key and relatively unique issues that emerged in our work together. The link between the subject areas we will touch on is that they function as the connective tissue weaving many fields of research together. Collectively, they afford an opportunity to propose and outline a comprehensive framework for ensuring the safety and ethical development of AI and robotics.

AI Safety and Ethics

Work on the safety of AI systems is a young discipline. For decades, researchers in AI have been focused on the challenge of building systems that perform discrete tasks. This was seldom easy. Paradoxically the pace of development has been rapid and slow. There have been constant discoveries and the creation of new tools, such as sensors and new algorithms, that quickly expand possibilities and approaches. Simultaneously, engineers have learned the difficulties involved in building systems to perform even basic tasks.

AI safety and ethics was largely a forward-thinking field inhabited by a small cadre of engineers and social scientists. The advent of deep learning and progress with other machine learning (ML) approaches led to the rapid development and deployment of AI systems. In turn, a large and growing community of researchers began working to ensure AI safety and to tackle ethics concerns that arise in the development and deployment of AI systems.

Some basic tasks have been particularly perplexing. For example, the founders of AI, who met at Dartmouth University (then Dartmouth College) in the summer of 1956, believed that the problem of computer vision could be solved by one graduate working on the task over one summer. More than sixty years later, accurate and reliable computer vision has not be fully realized, even though tens of thousands of engineers have worked on various facets of the problem.

Researchers who had been working in the field for 20 or 30 years, such as Cornell University's Bart Selman, had begun to feel that "various forms of perceptions, such as computer vision and speech recognition" were "essentially unsolved (and possibly unsolvable) problems." That concerned vanished with the advent of the deep learning approach to data analysis. Rapid advances have and continue to be made in computer perception. Some facial recognition software programs have accuracy rates of better than 90% in matching a photograph against a large database of images. While the failure rate is high, this degree of accuracy is useful for some applications.

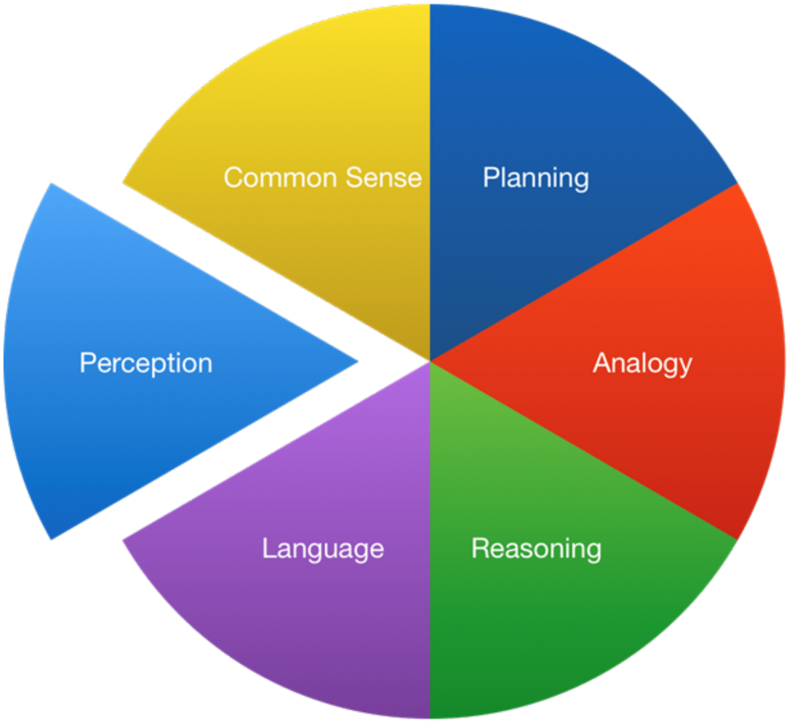
Deep learning emerging as a very useful AI approach was enabled by advances in computing power, hardware architecture, and the availability of large computerized databases dedicated to specific subject areas. Deep learning, in combination with other machine learning approaches, such as reinforcement learning and inverse reinforcement learning, has led to an explosion of useful applications, and many more applications lie on the near horizon.

More importantly, it made the development and deployment of AI applications relatively easy. Today, anyone with a large database of information on a single subject can probe that database using online tools to search for salient relationships within the data. This explosion in applications has been accompanied by concerns as to whether both the nearer-term systems and more advanced systems will be safe. Safety has always been important to engineers, but until the field grew robust, it did not receive much attention. In effect, AI safety was a necessary corrective in the trajectory of research that had, up to that point, concentrated on getting systems to function properly.

In February 2009, Selman and Eric Horvitz, then-President of the Association for the Advancement of Artificial Intelligence (AAAI), co-chaired the AAAI Presidential Panel on Long-Term AI Futures. At that gathering, certain speculative concerns, such as the possibility of smarter-than-human AI, were explicitly addressed, but they were not considered urgent or even likely to occur. "There was overall skepticism about the prospect of an intelligence explosion as well as of a 'coming singularity,' and also about the large-scale loss of control of intelligent systems," states the August 2009 interim report that summarized the meetings. A mere five years later, that assessment had radically changed. At the FLI workshop in January 2015, Bart Selman stated, "a majority of AI researchers now express some concern about superintelligence due to recent breakthroughs in computer perception and learning."

However, many within the AI research community were not happy that the birth of a rich AI safety trajectory had been promoted by a concern over the longer-term existential risks posed by the speculative creation of AGI. Underscoring the possibility of AGI contributed toward the public being fearful of AI. Wasn't it possible to enrich research on basic safety without making more speculative risks central to that research? We will return to this topic below.

AI will touch nearly every facet of modern life, from increasing productivity and efficiency to advances in healthcare. In a sense, AI is more similar to electricity than it is to a more sector specific technology, such as the automobile. Nevertheless, the tools and skill set available to contemporary machines is limited in comparison to that of humans. Deep learning has helped solve the problem of perception. Computers, however, are very poor at common sense, reasoning, planning, working with analogies, and understanding the semantic content of language.



Courtesy of Gary Marcus

Contemporary machine learning approaches are a breakthrough in AI. Should similar breakthroughs with far reaching potential occur in the near future, such as the ability to plan or reason, the safety and ethical challenges will also expand rapidly.

Section III. Silo-Busting, Collaboration, and Trust

If done well, workshops can create synergies, serve as catalyst to new insights, forge new relationships, expand awareness, and generate rich collaborative research projects to address areas of shared concern. A primary goal of the Hastings workshops was to expose the participants to fields of study that differed from but overlapped their areas of expertise. In addition to machine learning and other AI research, participants in the workshops represented social robotics, affective computing, cognitive science, machine ethics and robot ethics, engineering ethics, philosophy of technology, computer consciousness, decision theory, complex adaptive systems, resilience engineering, system testing and compliance, management of sociotechnical systems, risk assessment, and the law and governance of information technology and other emerging technologies.

Many of the fields of research represented at the workshops addressed challenges that would arise for ensuring safety, lowering risks, and minimizing undesirable societal consequences. Other fields would enhance the understanding of human ethics and hopefully facilitate aligning the actions of AI and intelligent robotic systems with human values. The working premise for the workshops posited that this exercise in intellectual silo busting would expand the understanding of challenges entailed in building intelligent, often autonomous systems, and foster collaborative projects. A subtext of this endeavor was the diffusion of any questions as to whether AI was being developed in a trustworthy and responsible manner.

In his 2014 book *The Innovators*, Walter Issacson emphasizes the importance of collaboration within fields of research and across disciplines, generations, and geographic boundaries in creating the digital revolution. “The main lesson to draw from the birth of computers,” Isaacson writes, “is that innovation is usually a group effort, involving collaboration between visionaries and engineers, and that creativity comes from drawing on many sources. Only in storybooks do inventions come like a thunderbolt, or a lightbulb popping out of the head of a lone individual in a basement or garret or garage. … what may seem like creative leaps—the Eureka moment—are actually the result of an evolutionary process that occurs when ideas, concepts, technologies, and engineering methods ripen together.”

AI researchers generally know each other. They often work together, adopt each other’s research for their own projects, and interact at annual meetings, including NIPS, the Conference on Neural Information Processing Systems, and those of their professional societies the Association for the Advancement of Artificial Intelligence (AAAI) and the International Joint Conference on Artificial Intelligence (ICJAI). In other words, collaboration among scientists is already central to the development of artificial intelligence and robotics within academic and corporate research laboratories.

As the field transitions to one whose societal impact is substantial, further collaborative relationships will be crucial for addressing the challenge of creating artificially intelligent agents that are beneficial, safe, and sensitive to human values in the choices and the actions that they take. The significant societal impacts of AI require that the community of collaborators goes beyond the AI research community and encompasses social scientists, policy planners, applied ethicists, and ultimately the public. At stake is whether the promised benefits of AI will be fully realized, whether there is sufficient trust of the researchers and industries developing and deploying the technologies, and whether there is sufficient confidence that the technologies being deployed are worthy of being trusted.

Many of the challenges generated by AI are technical in nature, and will be solved by AI researchers. Other challenges, such as AI's impact on employment, are not AI problems *per se* and will need to be solved by social theorist and policy makers. The more difficult issues are those where technical concerns and societal impact are entangled.

A tiny cadre of AI researchers formed ethics committees that met during professional conference, but until recently they had little impact. AI researchers only occasionally knew those working in complementary fields centered on the safety of computational and complex adaptive systems. Some of these other fields are broad in scope, such as engineering ethics, which applies ethical principles to the practice of engineering, and educates engineers to be sensitive to the ethical challenges that arise in their profession. Other research fields consider why systems fail or engage in forensic analysis of serious accidents such as airplane disasters. These investigations can reveal not only technical failures, but also inherent difficulties in coordinating the activities of humans and of autonomous systems such as airplane pilots, autopilots, and traffic controllers.

AI researchers function within a faith that their field will eventually reproduce, or at least computationally simulate, all higher-order mental faculties. They draw upon the ideas of cognitive scientists in developing models for their systems while not uncommonly brushing aside critiques as to whether consciousness or semantic understanding, for example, can be easily realized in silicon and software. Philosophers have played a unique role as both collaborators and critics in the development of smart systems. Often, that role goes either unacknowledged or embraced only after the fact, when the reservations of a philosopher, such as Hubert Dreyfus on *What Computers Can't Do (1972),* turn out to be more accurate than those of AI enthusiasts. Whether Dreyfus was correct as to whether there is anything computers will not eventually do, he was correct in pointing out naive assumptions that optimist researchers were making about the ease with which high-level cognitive capabilities would be realized computationally.

System failures are not always predictable or traceable to design or technical errors. AI systems, particularly those that will be deployed in the commerce of daily life, are complex adaptive or complex socio-technical systems, engaged in a complicated dance in which the technical artifact must smoothly integrate and adapt to the environment in which it operates. The behavior of such systems can at times be unpredictable or unanticipated. "Normal accidents" (see the book *Normal Accidents,* by Charles Perrow)*,* also known as "system accidents," occasionally occur, even when no one has done anything wrong. Recovering from failures with minimal damage requires engineering robustness and resilience into the systems.

Leading scholars in philosophy of technology and applied ethics, engineering ethics, cognitive science, systems theory and resilience engineering, risk analysis, decision theory, and the testing and management of sociotechnical systems may know each other. They may or may not, however, have a depth of understanding of the latest research, such as that in AI, and therefore are not necessarily prepared to evaluate whether enthusiastic expectation of significant breakthroughs is warranted.

The first challenge in bringing experts together from many different fields into a Hastings workshop was providing enough background so that each understood the work of the others present. One limitation of cross-disciplinary workshops is that only key ideas from each field can be shared, and participants do not become instant experts in new subject areas with which they are unfamiliar. However, they do acquire a rudimentary understanding of what each field has to offer, and hopefully draw upon each other when they should.

While each workshop included leaders in AI technical research, we were limited in how deeply we could probe more technical issues. This was a function of the difficulty in getting those without technical expertise to fully understand the technical problems, but also of the recognition that many issues in the technical design of AI algorithms were already being pursued by specialized teams. Nevertheless, the participants without an AI technical background were knowledgeable scientifically and anxious to understand as much as possible about research in AI. The focus of our workshops, however, sat on the cusp of concerns where technical challenges and societal concerns converge. While AI researchers might be prone to try to solve all issues through technical means, they were also coming to appreciate that some issues would force them to draw on the expertise of social scientists, the understanding of applied ethicists, and the experience in testing socio-technical systems and analyzing system failures.

The prospects for collaboration between AI researchers and those within other scholarly communities was the objective of our workshops. The subtext of these workshops, however, was trust, understood both in terms of the trustworthiness of the systems being developed and the trustworthiness of those developing and deploying AI systems. Trust is a multi-dimensional concern and is discussed more fully in the two essays by workshop participants attached below.

The good news is that the current generation of scientists is seriously concerned about the safety, controllability, and societal impact of the AI applications they develop and help deploy. Unlike past generations, where most scientists dismissed the societal impact of their discoveries and innovations as not being their problem, leading AI researchers acknowledge some responsibility for the tools and techniques they are developing.

The emergence of a community of transdisciplinary scholars overseeing the development, safety, and ethics of AI applicants is also good news. Leading organizations, such as the IEEE and the Partnership on AI, are helping solidify this transdisciplinary community of researchers, corporate leaders, and applied ethicists. The effectiveness of these transdisciplinary dialogues, however, has yet to be tested. Will collaborative projects to address shared concerns be developed? Will the social scientists work to understand specific research well enough to be of assistance to experts? Or will the AI researchers dismiss the social scientists and ethicists as unhelpful?

Of equal importance is the cultivation of both polymaths and a cadre of transdisciplinary scholars. These scholars may not always have expertise in the breadth of fields their interests encompass, but they nevertheless feel comfortable engaging topics that transverse many subject areas. Transdisciplinary scholars at their best recognize the limits of their understanding, but they are always open to learn and probe new realms of concern.

Unfortunately, our educational systems train specialists, and while colleges and universities give lip service to the need for interdisciplinary scholarship, few individuals are rewarded for such work. Interdisciplinary research is often left to older scholars who have already established reputations or obtained tenure for specialized research. Incentives are needed to cultivate a robust cadre of polymaths and transdisciplinary researchers. In our workshops, we addressed this lack symbolically by inviting a few young promising researchers, who had not yet fully established their reputations, but were seen as potential leaders in interdisciplinary research.

Today, industry leaders have begun to recognize that they need applied ethicists and others sensitive to the social impacts of applications that utilize digital technologies. But they are having difficulty finding individuals with sufficient expertise. The few older scholars with such expertise are insufficient to fill the demand.

Given the importance of collaborative—and often interdisciplinary or transdisciplinary—projects to high-tech innovation, it behooves us to create opportunities for scientists, engineers, philosophers, and social theorists to learn from each. We should not depend on serendipitous encounters. One way to facilitate this collaboration is through workshops that bring experts from many disciplines together to learn from one another. At the very least, workshops provide participant with a better appreciation for the richness of the challenge at hand. Workshops, however, are not sufficient. There is an increasing need for transdisciplinary research centers, technology review boards (Section 8), and other mechanisms for ensuring that collaboration between AI researchers, social scientists, and applied ethicists on the safety and societal impact of AI systems is ongoing.

Recommendation: 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.

AI: Whom Do You Trust?

David Roscoe

It is commonly perceived that the gating factor driving AI's transformation of society and the global economy will be the speed of technical innovation. In our series of workshops over the past two years, experts from several diverse academic, technical, and professional disciplines have collaborated to explore the ethical and social landscape being shaped in this transformation.

Throughout these discussions, as the sole layperson in the room, I observed discernible tension between those at the center of the AI innovation ecosystem—tech executives, computer and data scientists, engineers, and academicians in the technical disciplines--and those outside of the ecosystem who are concerned with its implications, including bioethicists, lawyers, economists, psychologists, others in the liberal arts disciplines, and public and civic advocates. Participants educated and challenged one another respectfully and without hostility, expressing a shared hope that genuine give-and-take collaboration would provide meaningful opportunities for outsiders to influence the arc of AI progress. However, there was also a nagging concern that collaboration might fall short, essentially leaving outsiders with little choice other than to trust those "in the know" to make all the right decisions for society. How trust develops between AI's insiders and the outsiders who represent wider society will have far greater impact on AI's transformation of society than the pace of technical innovation.

In the current AI landscape, questions of trust arise in three categories. First and most immediately, there are concerns about AI's first-order operational effects on consumers and users. Next, there are second-order, deeper societal questions about whether corporations and governments with enormous power to deploy AI as they choose have public and citizen interests sufficiently in mind. Finally, there are questions unique to the nature of AI itself, about its potential to evolve trust-based relationships with humans as collaborative advisors, assistants, servants, and partners.

Creating Consumer Trust in AI Systems: First Order Concerns

The AI community is focusing significant talent and energy to address the first-order effects of its current and planned AI products. Are they safe to use, or at least safe enough? Is the data upon which AI applications are being trained fair and unbiased, or is there unintended discrimination against minority populations? Are AI decisions transparent to humans, and can the rationales upon which they are based be understandable by humans?

These issues are being hotly debated in a broad range of areas: in transportation (assuring safety in self-driving vehicles with Level 3, 4, and 5 autonomy); in the justice system (eliminating bias in setting bail and sentences); in finance (eliminating racial bias in lending decisions); and in human resources (avoiding discrimination in corporate hiring, compensation, and promotion decisions).

There was ready agreement among all participants that a bottom-up approach led by the AI community itself to address these issues would be far more effective than immediate attempts to enact rigid laws or impose slow-changing regulations. To ensure trustworthy products, standards bodies including the IEEE and AAAI are working to create industry standard to institute proper safeguards around safety, bias, and transparency. Meanwhile, the *Partnership on AI For the Benefit of Humans and Society* (PAI), established in 2017, envisages itself as a collaborative platform within the AI ecosystem to develop corporate practices to address these same issues. Six major US technology companies (and, as of October, one major Chinese company) have invited over 80 civic and advocacy organizations from the US, Europe, and Asia to join in these collaborations.

These initial steps to adopt sound industry standards and to promote good corporate practice are showing encouraging progress. However, it is early, and these collaborative efforts are occurring in an intense competitive environment with enormous wealth and power at stake. There is still a degree of mistrust, both among the competitors themselves and between powerful companies, consumers, and civil advocates. Closing these trust gaps will occur in small steps not great leaps, and building on progress will require sustained leadership.

Creating Public Trust in AI Systems: Second Order Concerns

The deployment of AI across all domains of society also raises profound social and ethical issues well beyond the direct impact on consumers:

- Security and privacy: AI-empowered surveillance tools have the potential to drastically reduce crime, strengthen domestic security and improve public safety, but at what cost to individual rights to privacy? How much control should individuals have over the control and use of social media data?
- Warfare. Should autonomous lethal weapons be banned in wars on land, sea, air and space? How would their use alter both the nature of war and the prospects for future wars? Can AI-fueled cyber warfare, a fifth battlefront among nations, be contained as another cold war, and should it become hot, will cyberwar spread to other battlefronts?
- Fairness. Will the wealth generated by AI widen or narrow economic inequalities, both within and across national economies? Will AI products that improve the quality of life be accessible across the socioeconomic strata, or only to those who can afford them?
- Labor. Will AI create more jobs in the global economy than it decimates? And will this transformation occur in a few giant waves, or over a longer period? Should society prepare again for another new economy, or instead, for a new society? Will a new social contract be required, based on the reality that full employment is no longer a valid presumption?
- Democracy. The misuse of AI as a weapon shows signs of eroding public trust in free speech, a free press, and the election process. What responsibilities do AI and social media platform companies have to moderate AI-generated content that promote violence and hate?

These second-order issues cannot be fully resolved solely by the adoption of sound corporate practices and industry standards, although the AI community does have an opportunity and a duty to engage with policy makers and governments. Unfortunately, because AI technologies are spreading so rapidly and unpredictably, the knowledge gap between the AI community and the policy-making world is large and growing.

In the US, the disturbing growth of general mistrust between industry, government, and the citizenry has extended into the AI landscape. Recent missteps by major social media companies have increased mistrust of corporate motives by both government and the public. Some tech employees oppose cooperation by their companies on the development of AI-based surveillance and military capabilities. Minority and vulnerable populations feel underrepresented in the design of AI products and services. Unless these trends are reversed, this growing aura of mistrust will significantly hamper the development of sound US policies on privacy, security, fairness, labor, and democracy.

Countries and national cultures differ significantly on basic values and norms, including concepts of fairness and equity, and the relative priority of individual rights and community goals. These differences represent additional barriers to the forging of trust needed to achieve international standards and agreements.

Bridging, and then closing, national and international trust gaps between the AI community, the citizenry, and the policy world is essential for success. The first step must be to develop a shared awareness of what is actually happening and to foster mutual understandings of what is at stake. A major recommendation of this project is to institute a “global soft governance” process to accomplish precisely these goals. That process is more fully described elsewhere in this final report, but in this context, it should be seen as essentially a powerful international trust-building initiative.

Creating Trust in the Human-AI Relationship

AI systems, services, and products are being designed to assist humans, by making the tasks they perform less error-prone and more productive**,** to augment humans, by performing tasks humans are unable to perform on their own**,** and to replace humans, by relieving them of tasks when it is unequivocally clear that AI is able to perform them more cheaply, safely, or productively

This pursuit of human leverage and increased productivity has created a complex set of new dynamics between AI and humans, again rooted in trust. AI applications are being deployed as assistants, caregivers, and companions to humans by “learning” the habits, desires, and preferences of human masters. In these relationships, humans experience attention, care, companionship, friendship (and even love) based on a one-way trust of AI catalyzed by emotional responses to AI’s human-like expressions and body language. But are humans who experience such emotions genuinely benefited, or are they merely victims of emotional manipulation? Because AI today is essentially sociopathic, its rapid deployment is controversial in the AI community. Until there is a proven solution to the “value alignment problem”, discussed elsewhere in this report, some argue for curtailment of the use of AI that emotionally manipulates humans.

The Need for Trustworthy AI

Carla Gomes and Bart Selman

The transition of artificial intelligence (AI) from a largely academic endeavor to a discipline with significant societal impact has given rise to a range of new challenges. In particular, the AI research and development communities are developing new mechanisms to guarantee fairness, accountability, and transparency for the use of AI systems. We will argue that this endeavor requires a transdisciplinary approach, bringing in perspectives from a range of disciplines. Moreover, AI decision support systems will need to shift focus from the traditional single-objective optimization point of view to a multi-objective methodology. We will see how such multi-objective optimization enables users to consider various important tradeoffs in design objectives that AI decision support systems can reveal. AI systems will be able to consider such tradeoffs because they will be able to explore very high-dimensional spaces that are far beyond human cognitive abilities. Such super-human cognitive abilities expose some of the foundational limitations on the transparency of AI methods. As a consequence, we will argue that the adaptation of such AI systems in society will require the introduction of a level of multifaceted "trust" in such systems. We will discuss possible avenues for developing such trust in powerful autonomous AI.

Multi-Objective Criteria Require Transdisciplinary Research

There is a significant effort in developing techniques for obtaining so-called value aligned AI systems, in which human values and priorities are aligned with the built-in objectives of the AI systems. Even assuming significant progress can be made in this area, users of such systems still need to trust that the developers did consider a range of relevant factors such as those concerning ethics, safety, economic impact, transparency, etc. Such "trust in the design" can only arise from a close collaboration with other disciplines. We have seen this in our own work on Computational Sustainability, a field where computer scientists develop new computational methods to address the core environmental, economic, and sustainability challenges of our time (Gomes 2009). Progress in this area requires balancing economic, environmental, and societal issues. In many sustainability problems, it is therefore critical to jointly consider multiple, often conflicting, objectives, giving rise to challenges in multiobjective learning and optimization. For example, the United Nations lists 17 different sustainable development goals to consider for a sustainable world (United Nations 2015, 2017). We work with researchers from a range of disciplines, including ecologists, biologists, economists, civil engineers, and sustainability scientists. This is the only way to provide validity to our proposed solutions and models.

AI techniques are well-suited to analyze multi-objective forms of learning and optimization. AI methods can analyze the tradeoffs between criteria in these high-dimensional spaces and provide human experts with summarized visualizations of the tradeoffs. This work is a great example in which AI can complement humans by providing a higher-dimensional analysis, much beyond what is feasible for humans, but also lets the human experts explore the high-dimensional solution space through interactive visualizations.

A concrete example arises in our work on the analysis of the impact of the proposed construction of hydropower dams in the Amazon basin, with about 300 new hydropower dams proposed, which will dramatically affect a variety of Amazon ecosystem services, such as biodiversity, sediment transport, freshwater fisheries, navigation, besides energy production. Multi-objective optimization can identify the so-called Pareto frontier, which captures the trade-offs between the multiple objectives with respect to the different non-dominated solutions. The non-dominated solutions also provide valuable information concerning the dams' ranking. We have developed exact dynamic-programming algorithms, fully polynomial time approximation schemes, and other approaches for computing the Pareto frontier for tree-structured networks (Wu et al. 2018). For example, we can now approximate the Pareto frontier for the entire Amazon basin (approximately 5 million river segments), with respect to four criteria (energy, river connectivity, a good proxy for fish migrations and navigation, sediment production, and seismic risk) within 5% of the true optimal Pareto frontier in a reasonable amount of time. The results, combined with visualization tools, help policymakers make more informed decisions concerning multiple criteria and different planning geographic scales. Nevertheless, there are many more dimensions to consider. The next generation of AI systems will be able to learn and reason in much higher-dimensional spaces. These capabilities will lead to decision support systems that can manage much more complex tradeoffs than humans can consider, enabling better policy decisions.

Given the inherent limitations of human cognition, it will be infeasible for us to fully understand the high-dimensional analysis performed by AI systems. The systems will only be able to provide partial explanations concerning certain individual tradeoffs between criteria that guide the overall decision making. To deploy these systems, we will need to create a level of trust around the design and the performance of the systems. We will return to this issue below.

As a final example from computational sustainability, we consider a key issue in environmental policy which is the need to balance individual interests and the common good (Hardin 1968). In this area, game-theory models can model the interactions of multiple agents and show the effects of competing interests. In the context of natural resources or climate change on the international level, for example, economic incentives may influence whether a country is motivated to enter an agreement and then abide by it. Incentive-based policies can also facilitate sustainability challenges on a smaller scale (e.g., the establishment of novel markets for land-conservation activities). AI systems will provide powerful techniques for developing and analyzing such complex multi-agent multi-objective spaces, enabling better decision making and policy development. But, again, we will need to develop a framework within which policymakers can trust such AI design tools that operate beyond human cognition.

Inherent Limits to Transparency and Interpretability of AI systems

There is a rapidly growing research effort to develop new mechanisms to guarantee fairness, accountability, and transparency for advanced AI systems. In particular, in the area of AI decision support systems that rely on sophisticated machine learning capabilities, questions of how to deal with inherent bias in the training data and the interpretability of the learned statistical models are receiving significant attention. Good progress has been made to address several of these issues. Nevertheless, we also need to consider the limits on our ability to address these issues.

As we discussed above, AI systems can provide powerful insights into high-dimensional spaces that lie beyond human understanding. The insights from such systems can actually lead to better decisions in many respects, so we need to find ways to make such systems acceptable for practical use. In order to do so, it is useful to first look at some distinctions concerning the performance and interpretability of AI systems. We will do this by taking a closer look at certain specific examples.

We first consider AI systems as they are being developed for self-driving cars. Such systems are part of a broader class of AI systems, called cyber-physical systems. One of the primary concerns with respect to self-driving cars is the notion of safety. Ideally, one would like to provide specific safety guarantees. This is an area of engineering that is fairly well understood. Safety guarantees can be provided in part empirically by test driving the cars in a supervised setting (i.e., with human co-driver present) but also by analyzing the systems using techniques from the formal verification of hybrid systems. As in general for engineering design, adding redundant safety systems can be used to reach almost any desired level of safety, at an additional expense of course. Given that the goal is not to eliminate all possible accidents but rather to significantly improve over a human driver level of safety, the goal of self-driving cars is within reach with current AI technologies. Moreover, since the systems can be validated with safety testing protocols that are already quite well-developed, it therefore appears likely that the technology will be accepted by the broader public. We should note that certain specific ethical issues such as those related to the trolley problem will still need further exploration. Nevertheless, such issues are unlikely to hold up the introduction of self-driving cars, since similar issues did not prevent the use of autopilots in other systems either. In fact, the widespread use of human-driven cars has shown that people are quite willing to accept a certain level of risk in these domains.

Self-driving car technology falls in the relatively well-understood domain of transportation and control engineering with the addition of new sensors in the form of vision and other sensory capabilities. We expect the introduction of a range of other AI systems that operate similarly in relatively controlled and well-understood settings. Examples would be cleaning robots or other robots for executing routine tasks. In these domains, aspects of safety, transparency, and trust will be manageable with good design and engineering practice. The picture changes dramatically, however, when we consider systems that operate significantly beyond human cognition, such as systems that analyze ill-structured high-dimensional spaces or operate in adversarial multi-agent settings. Such systems lead to more significant challenges. To understand these challenges it will be useful to bring in some insights from the formal theory of computational complexity.

Computer chess is a good domain to illustrate issues in super-human adversarial reasoning. We currently have two different approaches for playing chess at a super-human level. One is a search-based approach, as exemplified by IBM's Deep Blue program that defeated the human world chess champion, Garry Kasparov, in 1997 (Deep Blue 1997). The other is a recent deep learning based program, called AlphaZero, that trained a deep neural net via pure self-play to become, in 2017, the strongest chess playing program in history (Knight 2017). (AlphaZero was designed for chess as follow-up on the deep learning breakthrough for Go, called AlphaGo.)

A basic question we can ask of these two examples of super-human AI systems is “Can a human understand the chess moves made by either Deep Blue or AlphaZero?” However, given the complexity of chess, it seems unlikely that humans can fully understand the moves suggested by either system. More broadly speaking, we can ask whether we do have a guarantee that a move suggested by Deep Blue or AlphaZero is better than other possible moves at a given position (i.e., can we “trust” the systems to make optimal or near-optimal moves). The answer to this question is that we have no such guarantees either. Moreover, given our current understanding of the chess search space and issues of computational complexity, we have reasons to believe that we may never obtain such guarantees. So, any formal guarantees appear out of reach. Without such explicit guarantees, this leaves the question of whether there are other arguments that enable us to trust these AI systems to make the optimal or near-optimal moves in the domain of chess.

Interestingly, Deep Blue and AlphaZero are quite different with respect to the question as to whether they can be trusted. DeepBlue proceeds by exploring a very large set of possible lines of play, looking carefully many moves ahead, including all possible countermoves by its opponent up to a certain depth. Doing this type of search is far beyond any human capability. However, since we can understand the basic strategy (analyze all your possible moves, all possible countermoves, all possible next moves, etc., 15 to 20 or so moves ahead). If it was possible to carry out such an analysis till the end of the game, we would have a perfect chess playing program. Therefore, with significant depth of the move analysis, we can have confidence in such an analysis, which means we can have trust in the quality of Deep Blue’s play and move recommendations. This is an interesting example of how we can trust an AI system without the AI system being able to explain its particular actions to us.

The situation is different for AlphaZero. AlphaZero makes moves based on how its learned deep net judges the position. More specifically, its deep net judges whether a particular board is a board that is likely to lead to an ultimate win, loss, or draw. That is, in a very real sense, it tries to recognize a position as a potential “win”, “loss”, or “draw.” Recent work in computer vision has shown that deep nets for recognition can be fooled by cleverly constructed images that mislead the deep net recognizer. Similarly, it may be possible that AlphaZero also significantly misjudges certain positions. This means we cannot fully trust AlphaZero’s play. Of course, even if such misjudged positions exists, it may be hard for an opponent to reach them during actual play. Moreover, AlphaZero also incorporates a search component which will somewhat mitigate the effect of errors in the classification of positions. Nevertheless, AlphaZero is a good example of a AI system exhibiting super-human performance but with no guarantee that it cannot be fooled by some hidden clever play strategy. We would, therefore, argue that AlphaZero is actually in a sense less trustworthy than a search-based chess architecture, even though it is overall the stronger player! This issue arises in part because neither architecture provides a perfect player and neither is able to give an interpretable justification for its choice of moves.

The attentive reader may have noticed a possible hidden contradiction here. How can we have more trust in Deep Blue’s moves than in AlphaZero’s moves, while AlphaZero is the stronger player? The reason is that while overall AlphaZero is indeed the stronger player, it may have some hidden weaknesses that would not trip up Deep Blue. Specifically Deep Blue can guarantee that it will not lose within a certain number of moves (if feasible), because of its systematic lookahead strategy. In contrast, AlphaZero cannot provide such kind of guarantee. So, a strategy that would defeat AlphaZero, if it exists, would lead to a very unusual game that would only work against AlphaZero's deep net. Such a strategy would steer towards certain board positions that AlphaZero's net significantly misjudges. These board states are analogous to certain images that fool deep vision systems. Deep Blue would likely not be fooled by such positions because Deep Blue systematically analyzes any board position in an actual game to a significant depth. (Note of course that Deep Blue is not perfect; it cannot analyze to the full depth of the game tree and uses a heuristic board evaluation after looking ahead a significant number of moves.) So, we have two AIs of comparable strength but one is arguably more trustworthy than the other.

As a final example, let us briefly consider a recent result in automated reasoning for mathematical discovery. In 2014, the so-called Erdős discrepancy conjecture for a discrepancy of 2 was resolved using an automated reasoning system (Konev and Lisitsa 2014). The conjecture had been open for over 60 years. The proof involved showing that a certain type of sequences of +1s and -1s does not exist. To show the non-existence of a mathematical structure with a certain property generally requires enumerating all possible structures and showing that each one does not have the desired property. For the Erdős conjecture, the exhaustive enumeration of all sequences would require more compute power than is available on our planet running longer than the expected lifetime of our solar system. Fortunately, the AI reasoning engine discovered a much more clever way to analyze the space of sequences and was able to show in less than one hour on a MacBook Pro that no sequence with the desired property exists. Now, how can we trust this result? If an exhaustive enumeration was feasible, one could potentially verify the code for generating the sequences and formally prove the code correct. However, the AI reasoning system introduced a series of very clever search reduction steps ("shortcuts through the search space"). This requires highly complex code that is beyond any current verification procedures. However, remarkably, the system did also generate a special formal proof trace, stored on disk, that can be checked independently. Each step in the formal proof represents a very small logical inference step easily checkable by hand or with a short piece of Python or Java code. Simple making sure each step is correct validates the overall proof. This can be done in a few minutes using Python or Java code. The verifiable proof trace makes the result fully trustworthy. In this setting, we do have yet another form of trust. In particular, we don't need to trust the correctness of the original AI reasoning system, since the system did provide an explicit proof that can be checked separately by a very simple proof checking program. This is an example of trust through verification but with the added feature that the AI system itself provides the object (the proof) that captures the validity of the final result. So, the system generates its own verifiable explanation. This level of trust is fully acceptable for mathematicians who can now build further mathematics on top of the conjecture proved by an AI reasoning system. As an interesting aside, the proof trace is actually 13 GB long! This was the longest rigorous proof in mathematics at that time, and the first non-trivial automatically obtained mathematical result with a fully verifiable proof trace. Mathematical results of this form are likely to become an integral part of our mathematical world.

Conclusions

The development of trustworthy AI will require transdisciplinary research teams that consider the tradeoffs between criteria in multiobjective learning and optimization. We have argued that given the likely limits on providing true human understanding of the high-dimensional analyses performed by advanced AI, we need to develop a notion of trust in such systems in order for them to be successfully deployed. The development of trust in such AI systems may follow a path that is also used for human expertise. We are accustomed to trust the expertise provided by a variety of human experts such as medical professionals, civil engineers, legal professionals, etc. Society gains such trust over time in part through the guidance of professional organizations that consider multiple risk factors, ethics concerns and other issues relevant to the field of expertise. We may need similar organizations to develop advanced trustworthy AI systems.

Our discussion of Deep Blue vs. AlphaZero for Chess and the resolution of Erdős conjecture using an AI reasoning system show AI systems that provide different levels of trust to the user. However, because of fundamental results from computational complexity, it is likely that certain types of AI systems that explore high dimensional spaces can only be trusted based on careful empirical validation. In practice, we envision a whole variety of trust mechanisms for advanced AI systems to be developed depending on the criticality and nature of the application domain.

References

Deep Blue 1997. Deep, Deeper, Deepest Blue. *New York Times*, May 18, 1997.
https://www.nytimes.com/1997/05/18/weekinreview/deep-deeper-deepest-blue.html

Gomes, C. 2009. Computational sustainability: Computational methods for
a sustainable environment, economy, and society. *The Bridge* 39, 4 (2009), 5–13.

Hardin, G., 1968. The tragedy of the commons. *Science* 162(3859): 1243–1248.

Knight, W., 2017. "Alpha Zero's "Alien" Chess Shows the Power, and the Peculiarity, of AI". *MIT Technology Review*, 8 Dec. 2017.

Konev, B. and Lisitsa, A., 2014. A SAT attack on the Erdős discrepancy conjecture. In International conference on theory and applications of satisfiability testing (SAT-14).

Silver D., Hubert T., Schrittwieser J., Antonoglou I., Lai M., Guez A., Lanctot M., Sifre L., Kumaran D., Graepel T., and Lillicrap T., 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv preprint arXiv:1712.01815. 2017 Dec 5.

United Nations, 2015. Transforming our world: The 2030 Agenda for Sustainable Development (2015).
http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/&Lang=E

United Nations, 2017. The Sustainable Development Goals Report. 2017
https://unstats.un.org/sdgs/files/report/2017/TheSustainableDevelopmentGoalsReport2017.pdf

Wu, X., Gomes-Selman, J., Shi, Q., Xue, Y., Garcia-Villacorta, R., Sethi, S., Steinschneider, S., Flecker, A., and Gomes, C.P.. 2018. Efficiently Approximating the Pareto Frontier: Hydropower Dam Placement in the Amazon Basin. In Proc. AAAI-18, 2018.

Section IV. Addressing Nearer-Term Ethical Challenges versus Longer-Term Approaches to AI Safety

Three tensions pervaded our conversations:

1. A significant portion of the attendees were focused primarily on ameliorating any harms from future artificial general intelligence (AGI) or artificial superintelligence (ASI). The rest of the participants considered the advent of AGI either highly speculative, unlikely, or not warranting much consideration over the next 10 to 20 years. Furthermore, a few participants felt that dwelling on speculative future possibilities created unnecessary fear in the public that might stultify support for a field that offers untold benefits.

2. While most of the participants were concerned with minimizing the risks and undesirable societal impacts of nearer-term AI hardware and software agents, a few members concerned about AGI doubted that this would be of any help in ameliorating existential risks posed by AGI. 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?

3. AI researchers championed a value alignment approach to building sensitivity to human values into machine learning systems, while participants with a background in ethics, and specifically machine ethics, felt that the value alignment approach as initially defined and described lacked rigor and was unlikely to be successful.

Debate as to whether artificial general intelligence is realizable, will be created in the next 20 to 100 years, and will be beneficial or pose an existential risk to humanity is ongoing. Differences in opinion hinge partially on definitions of intelligence and differing understandings about the prerequisites for machines whose intelligence is comparable to or exceed that of humans. But there are also competing visions of what can be realized through the digital and computational mechanisms that are foundational to present-day systems. We will not attempt to resolve those debates here, nor are they necessarily resolvable given present-day knowledge.

Skepticism over the realizability of AGI is not disappearing any time soon. Nevertheless, Stuart Russell captured a sentiment that brought consensus to a Hastings workshop when he asked: "Even if there were only a ten percent chance that superintelligence is developed in the next fifty years, wouldn't you want us to begin work to ensure its controllability or safety now?" Everyone in that workshop agreed that the possibility of superintelligence should not be dismissed. Even if the odds were low in the estimation of some experts, they were high enough that the prospect should be taken seriously.

This, however, begs important questions, including: how seriously should those prospects be taken, how much investment should be made in research specifically directed at the guaranteed beneficiality or controllability of AGI, and can the safety of AGI be shaped by research on nearer-term societal and ethical concerns. Considerable funding is already directed at ensuring the safety of AGI. How much more funding is required is subject for debate.

Proponents of research directed specifically at ASI systems argue that it will be of a totally different order than the systems we are creating presently. Its intelligence will be so great that it will have the capacity to work around and defuse any safety mechanisms built-in that thwart achievement of its goals. Two basic conclusions follow from this contention: 1) Research on building sensitivity to ethical considerations in nearer-term systems is not central to ensuring the beneficiality and safety or ASI, and 2) the only way to guarantee that ASI does not pose an existential threat to humans is to align its values and goals with that of humans.

One critique of this position is that any research on ASI presumes that we can truly conjecture what or how ASI systems will emerge in the course of AI development. But the actual platforms upon which more sophisticated AI will evolve are likely to be determined by which solutions address nearer-term challenges. For example, there is considerable concern today that deep learning systems lack transparency—that is, that how the output of these systems is arrived at cannot be explained. If a substantial degree of transparency or explainability is not achievable for machine learning systems, then arguably the use of this technology for determining actions in mission critical contexts should be rejected. In other words, deep learning technology may or may not be central in the development of more sophisticated systems. Regardless, work on transparency and explainability for machine learning systems is getting serious attention and, if achieved, will have import for both nearer-term and longer-term applications.

In the evolution of the controllability or safety of AGI, the term "provably beneficial AI" emerged. This was due to a number of factors. Work on safety should be taken for granted as it is central to all engineering. Controlling AGI may not be an option—thus the focus on aligning its values and goals with those of human. More importantly, safety and controllability suggest that AGI, and by association AI research in general, is risky. This risk framing is likely to undermine support for research on AI. Therefore, it is important to underscore the benefits of AI research and that the responsible goal for any advances in AI should be that the technology's benefits are guaranteed, which entails that the risks have been defused.

The second meeting of leaders in the field of AI convened by The Future of Life Institute at Asilomar Beach, California on January 2017 issued 23 principles for the development of AI. A core Asilomar principle states: "Value Alignment: *Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation."* Value alignment has become an important concept among AI safety researchers. It is often referred to as a central problem that must be solved. However, at times "value alignment" refers specific approaches to solving this problem, as it does in the essay by Stuart Russell that follows.

When the term "value alignment" was first proposed, few in the AI community were aware that there was already a community of scholars who had created a discipline over the preceding decade known as machine ethics or machine morality. The philosophers, computer scientists, and practical ethicists developing machine ethics or moral machines were interested in developing computational systems with sensitivity to human ethical concerns, and capable of factoring these concerns into their choices and actions. The challenges these researchers worked on were sometimes rudimentary, such as what a robot delivering drugs to a home bound patient should do if the patient refused the medication. More broadly, however, machine morality considered ways of building ethical subroutines into systems that would guide the selection of choices in ethically significant situations towards existing norms and appropriate behavior. While machine ethicists considered challenges in scaling the moral intelligence of advanced systems capable of being declared the equivalent of human level moral agents, this was secondary to taking first steps toward forging methods for imbuing computational systems with very basic ethical decision-making routines.

Within the machine ethics community, moral philosophers analyzed whether ethical principles and theories, such as the Ten Commandments, utilitarianism, Kant's categorical imperative, and the Principles of Biomedical Ethics are computationally tractable. Whether Asimov's three laws for robots (later four, with the addition of a Zeroth law) are useful and feasible has been among the issues given attention. The field and initial approaches considered for machine ethics were mapped in the 2008 book *Moral Machines: Teaching Robots Right From Wrong,* by Wendell Wallach and Colin Allen.

In 1950, Alan Turing proposed a program of learning, similar to that given to a child, as a route for developing machine intelligence. In *Moral Machines,* Wallach and Colin propose that a learning system directed at acquiring character traits (virtues) might offer the best pathway for ensuring that sensitivity to value/moral considerations is deeply ingrained and would be stable under fire. They characterized a system that learns about values, ethics, and mores as a bottom-up approach to machine ethics. However, machine learning algorithms were certainly not up to the task in 2008.

Value alignment can be thought of as a bottom-up approach to the development of moral machines. A central inspiration for the Hastings workshop was to bring AI researchers proposing approaches to value alignment together with leading members of the machine ethics discipline to learn from each other. Our hope was that they could collaboratively explore whether breakthroughs in machine learning opened the door for a truly feasible bottom-up approach to the realization of moral machines.

The AI researchers were open but skeptical that ethicists had much to offer in the development of value aligned AI. The machine ethicists considered the approaches outlined by AI researchers as somewhat naïve. AI researchers tended to feel that ethicists made the issues too complicated. Practical ethicists viewed the approaches offered by AI researchers as overly simplistic. This tension and its evolution toward a degree of mutual understanding of the shared goal is outlined in the essay below by Shannon Vallor.

The tension between machine ethics approaches and value alignment to ensuring the safety of AI systems remained a driving force throughout the three workshops, as did discussing the importance of focusing on nearer-term challenges for addressing AGI concerns. During the first workshop, key participants introduced their approaches. The second workshop dove into value alignment and machine ethics approaches in greater depth.

During the third workshop, we used a case study to explore the evolution of AI toward greater intelligence, and the ethical challenges arising in that process. The case study considered the next stages in the evolution of personal digital assistants (PDAs) built into smart phones and other devices. PDAs have already been used to autonomously make reservations at restaurants or purchase airline tickets. What troubles might PDAs cause as they learn all of one's preferences and attempt to fulfill one's goals and desires autonomously? How can we ensure the actions of a PDA will not be illegal or immoral? A further stage of development for a PDA will be collaborating with other PDAs, at first to schedule meetings when all parties are available. Unfortunately, one can also imagine PDAs coordinating stock purchases or performing other mutually beneficial tasks that border on illegal forms of collusion for the manipulation of markets.

Whether the approach is called value alignment or machine ethics, the challenge of imbuing AI systems with sensitivity to human values is an exciting project and central to the evolution of beneficial and safe AI. Thus it forms a second half of the recommendation introduced in the last section of this report.

Recommendation: 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.

Value Alignment and Machine Ethics

Shannon Vallor

Overview

Implementing sensitivity to norms, laws, and human values in computational systems has transitioned from philosophical reflection to an actual engineering challenge. A 'value alignment' approach has gained traction with many AI researchers as a potential way to meet this challenge. ‘Value Alignment,’ which seeks new computational techniques for keeping AI system goals reliably aligned with the values of human users, may be seen as serving dual purposes. In the near-term, it aims to ensure the safe, ethical, and trustworthy performance of existing AI systems that have only ‘narrow’ or task-specific intelligence, while in the long term, value alignment is proposed as a way to ensure the safety of more advanced and hypothetical forms of artificial intelligence, including artificial general intelligence (AGI) or possibly AI that exceeds human intelligence (‘superintelligence’).

However, value alignment is not the only methodological approach to meeting the challenge of building safe, ethical and reliable AI systems, nor the earliest. Within the field of applied technology ethics developed by philosophers and other scholars trained in moral theory, ‘machine ethics’ has long been proposed as a path to safe, trustworthy, and ethically reliable AI performance (Wallach and Allen 2009). Machine ethics is a field of research compatible with a variety of potential engineering approaches, and need not be incompatible with many techniques favored within the value alignment strategy. However, machine ethics frames the task of building safe and ethically reliable AI systems in a fundamentally different way than value alignment proponents, in part due to different disciplinary orientations to the challenge.

Each of the Hastings Center workshops on *Control and Responsible Innovation in the Development of AI and Robotics* fostered extensive discussion and exchange among advocates of both value alignment and machine ethics, as well as the input of experts in law, policy, and social science whose insights illuminated the debate. In this section we summarize the key points of that debate, including the areas of greatest consensus and potential collaboration, and the issues upon which value alignment and machine ethics advocates remain divided.

The Value Alignment Approach

The value alignment strategy posits that values can be learned by observing human behavior. Its defenders often eschew the languages and principles of normative ethics in favor of more computationally friendly concepts, such as utility and reward functions, system goals, agent preferences, and value optimizers. Unlike normative concepts of justice, benevolence, duty, and virtue, these conceptual tools of the computational value alignment approach carry no intrinsic ethical significance; they could be used to describe amoral or even immoral agent behavior. However, many supporters of value alignment present their approach as simply a practical translation of utilitarian ethics: that is, a mechanical path to an ideally rational and ethical decision calculus by means of a machine learning method for understanding—and remaining behaviorally aligned with—individual and/or aggregate human preferences.

Vivid examples that express the need for a value alignment strategy abound. Some are fairly silly, such as Nick Bostrom's notorious 2009 thought experiment involving an AI tasked with maximizing a paper clip factory owner's goal of making as many paper clips as possible. Bostrom asks us to envision the AI becoming intelligent enough to become a devastating 'paper clip maximizer' on a mission to convert all the planet's resources into paper clips.
As absurd as that may seem (how exactly does the AI gain access to all the planet's resources?), consider the more plausible example of the artificially intelligent robot barista tasked with learning ever more creative and efficient ways to get coffee to the shop's customers more quickly and efficiently. All goes well until the robot discovers the most efficient path to the goal of all—throwing the hot coffee directly at the customers' faces.

The point of such examples is that even if we set aside the problem of translating human goals from natural language into programming languages, any human goal we can specify to an AI ('make a lot of paper clips;' 'get hot coffee to the customer quickly') will still be a narrow proxy for the richer, more contextually appropriate goal that includes all manner of unspecified values, assumptions, conditionals, and restrictions that an intelligent human would understand implicitly. Add to that understanding gap the ability of machine learning to enable exploration of novel, undemonstrated strategies (as with AlphaGo's creative play), *then* couple that with access to real-world kinetic power (through robotic actuators or autonomous software control of physical systems), and you have a recipe for disaster (Omohundro 2008, Bostrom 2009). Thus the need for AI safety research as a corrective complement to research focused solely upon the functionality of AI systems.

Within AI safety research, "value alignment" was proposed by Stuart Russell and others as a means to ensure that the values embodied in the choices and actions of AI systems remain in line with those of the people they serve (Russell, Dewey, and Tegmark 2015, Taylor et al. 2016). Value alignment quickly caught on within the AI safety research community. A core concern for many of the AI safety researchers attracted to value alignment is the need to ensure that any future artificial general intelligence (AGI) or superintelligence would be friendly to human values and aligned with human interests, survival, and needs. But as our robot barista example shows, even narrowly task-driven machine learning systems of the sort we can build today could benefit from advances in value alignment techniques. AI researchers working on value alignment have thus begun to direct attention to ensuring that systems fulfill nearer-term tasks in an appropriate manner. Nevertheless, value alignment as a research trajectory has remained concerned with laying foundations for an approach to values that can be scaled up to guarantee the safety and human-friendly behavior of AGI systems.

Machine Ethics

Predating the value alignment research agenda, but emerging from a very different disciplinary orientation, is the field of *machine ethics* (sometimes known as *machine morality, roboethics,* or *computational ethics*). At the *Control and Responsible Innovation* workshops, it became clear that value alignment and machine ethics proponents were not always well aware of one another's research efforts. Disciplinary siloing (see Section IV) in computer science, engineering, philosophy, and other related fields has hampered cooperation, knowledge-sharing, and collaboration to date among these researchers. During the workshops, the foundations of such cooperation were laid in part through clearer articulation of the two research agendas and their differences, which allowed for substantive mutual questioning and critique of the conceptual, methodological, and practical differences between the two. Those differences are sketched below, along with potential areas of future cooperation and collaboration between the research bodies.

The central topic of machine ethics is the theoretical and practical prospects for building moral machines. This goes beyond the goal of building safe and beneficial to humans, machines whose goals remain reliably aligned with human values. For as long as that goal were satisfied, value-aligned machines could remain entirely *amoral* in their own reasoning. In contrast (although there can be some overlap between different versions of these aims), the goal of machine ethics would be to build a system that *itself* possesses some degree of ethical intelligence and decision-making ability (Moor 2006). Such a machine would also be safe and beneficial to humans in its actions; but this would be enabled by some mechanical capacity for moral reasoning (or an artificial emulation of it). Thus the addition of the moral reasoning capacity is not of purely philosophical interest; rather, many defenders of machine ethics hold that imbuing machines with some form of engineered moral intelligence may well be the only reliable means by which safe and beneficial AI systems could be attained.

For many philosophers considering the prospect of imbuing computational systems with ethical behavior, machine ethics is a largely theoretical challenge. But some interdisciplinary teams have begun work on computational pathways for implementing moral decision-making capabilities in machine systems. The techniques they utilize are typically not the machine learning algorithms increasingly favored by AI researchers, but ‘top-down’ methodologies of constraint by deontic moral logics, decision trees, and so on (Anderson and Anderson 2015; Govindarajulu and Bringsjord 2015, Bringsjord et al. 2018). Such mechanisms primarily depend upon rule-driven directives (‘thou shalts and shalt nots’) to encode moral intelligence.

However, a purely top-down strategy is likely to result in brittle systems with limited abilities to hand complex or novel moral circumstances. Such systems may only be able to achieve what Allen, Smit, and Wallach (2005) call ‘operational morality.’ Operationally moral systems are those that function within boundedly moral contexts, in which the engineers and designers can discern in advance the array of challenges the machines will encounter. In effect, the computational system is programmed in advance to act appropriately in each situation it will encounter. When designers and engineers cannot predetermine all the circumstances an artificial agent will encounter, it becomes necessary for the agent to have subroutines that facilitate making explicit moral decisions. Even here, however, gaps in moral competence will remain, since the subroutines themselves cannot encapsulate all possible conflict resolution strategies that may be needed--especially if the system is allowed to operate ‘in the wild’ in unconstrained social environments. On this view, we should expect the best results from ‘hybrid’ approaches to machine ethics that merge top-down constraints with ‘bottom-up’ processes of machine learning that allow the system to gradually acquire flexible moral competence in the world (Wallach and Allen 2009, Abney 2012).

Over the coming decade or two, most artificially intelligent agents will continue to be single-purpose machines operating in boundedly moral contexts, and their explicit moral reasoning will be limited to determining which norms or courses of action apply in the situation at hand or when values conflict. For example, a caregiving robot attending to a homebound or elder person might have to select whether to deliver medicine on schedule or whether to stop and recharge its battery. The right course of action could depend on how critically the individual needs the specific medicine, or what might occur if the agent fails to recharge its battery immediately. Given limitations in the cognitive capabilities of present-day AI systems the contexts within which they can function appropriately are limited. However, as breakthroughs are made in machine learning, the environments within which intelligent systems can operate safely and even ethically will expand.

Machine Ethics and Value Alignment in Contention

The *Control and Responsible Innovation* workshops fostered discussion among machine ethicists and value alignment researchers who share the goal of building AI systems that are safe, reliable, and beneficial to humans. Yet the workshops also illuminated key conceptual, methodological, and practical differences in the ways this goal is sought:

1. *Conceptual Differences*
Value alignment approaches, as noted above, tend to eschew robustly normative concepts such as *justice*, *rights*, or the *good,* relying instead on formalisms expressed as utility functions and preference optimizers. Many ethicists regard these formalisms as inadequate proxies for ethical principles and concepts which they see as necessary to inform the proper expression of human values in machine code. Even the simplest core values of 'safe' and 'beneficial' AI are laden with many other implicitly normative concepts, an understanding of which is required for their appropriate interpretation. To give one obvious example, the intelligent robots in Isaac Asimov's stories have a dangerous tendency to misinterpret their goal of keeping humans 'safe' in a way that fails to grasp the importance of human freedom, dignity and autonomy for human welfare. Value alignment defenders respond that such examples are exactly the *point* of the value alignment strategy. This potential for machines to fail to accurately read and calculate human preferences is the very problem that the value alignment approach seeks to solve.

Ethicists, however, challenge the implicit assumption that concepts such as 'justice,' 'freedom,' 'dignity,' 'autonomy' can be adequately expressed in mathematical reward or utility functions that treat these as preferences to be optimized. For each of these moral values stands in perpetually fluctuating tension with many others, in ways that vary according to a practically infinite combination of contextual factors. For example, while it is true that human autonomy must often be respected even at the expense of human safety (otherwise any of us could be indefinitely confined against our wishes to a padded room), in other contexts we reasonably prioritize safety even over the wishes of rational adults (consider seat belt and helmet laws, or mandatory vaccination programs). There is no fixed rule that can tell a machine, or a human, how to make optimal or even acceptable tradeoffs among the relevant values, and yet we *do* make them, for reasons we can articulate and defend. Nor are these defenses naturally decomposable into statements about mere preferences, since neither the preferences we express at any given moment, nor the preferences we have expressed in our past behavior, reliably capture what we *ought* to do.

At this level, the disagreement between machine ethicists and value alignment defenders seems to be about how we should *talk about* or conceptually frame the problem, and which way of doing so is more likely to lead us to the right kind of engineering solution. Value alignment defenders regard the problem as one of teaching machines to properly read, weigh, and obey human ‘preferences’ in relationship to specific tasks and goals (all of which should, in principle, be specifiable in computational terms). Machine ethicists often regard the problem as one that will require embedding in machines some capacity to *make use* of explicitly and irreducibly normative concepts and principles of ethical decision-making, such as rights, duties, justice, the good, or virtue. Although this ability must be encoded by machine syntax, the code must *enable* ethical reasoning, not dispense with it. On this view, it is highly implausible that ethical concepts could be decomposed into a preference matrix without undermining the normative integrity and contextual sensitivity of their action-guiding structure.

2. *Methodological Differences*

These conceptual differences, then, translate into different views of the kind of engineering approaches that are likely to do best at developing safe, beneficial, and ethically reliable AI systems. Value-alignment approaches tend to favor techniques such as inverse reinforcement learning (IRL), a bottom-up form of machine learning in which the system infers its reward function from repeated observations of human actions (Russell, Dewey, and Tegmark 2015.) Of course, it will matter a great deal what sort of human serves as the subject of observation; given the ubiquity of human behavior that is neither safe, nor beneficial, nor ethical, this presents a significant methodological obstacle. A system learning to be reliably aligned with these values via IRL would need to be able to learn everything it needs from a narrow and highly controlled set of training data.

The question of *what the AI actually learns* also remains; from the standpoint of the ethicist, any behaviorist method of inferring human norms from observable actions, IRL included, will miss the moral content that constitutes and justifies those actions. To paraphrase Socrates’ lesson in the *Euthyphro*: the ethical agent’s actions are not moral because they are *done* by the ethical agent--they are *done* by the ethical agent *because they are the moral actions*. The internal considerations that lead the ethical agent to choose those actions are hardly inconsequential; they make it possible to effectively communicate our moral intentions, be accountable for outcomes, negotiate value conflicts with others, and learn from our moral mistakes (Arnold et al. 2017). An AI system that learns only the patterns of moral behavior can do none of these things, and will be severely hampered in moral performance as a result.

That said, bottom-up machine learning approaches such as IRL have considerable advantages over symbolic, rule-driven forms of moral logic programming, which suffer from the same problems of rigidity and contextual insensitivity as deontological ethics does in the human domain. For example, the ethical prohibition against lying in the Kantian ethical system has a multitude of apparently necessary exceptions (such as lying to the ‘inquiring murderer’ or the ‘inquiring death squad at the door’). This either means that the moral prohibition against lying does *not* hold universally (in which case we face the problem of being unable in advance to specify all the possible exceptions to it), or it does hold universally, but can conflict with other *prima facie* moral duties that are equally universal, in which case we need a scheme for prioritizing some universal rules and duties over others.

Unfortunately, there is no compelling universal rule that tells us how we should prioritize universal rules or rank conflicting ethical duties and values; the priorities and rankings that make moral sense and can be defended by reasonable person seem to vary according to the unique demands of the moral situation. An AI system programmed with deontic moral logic is either going to make potentially grave ethical errors due to its rigidity, be paralyzed by value and rule conflicts, or resolve such conflicts in unreliable and often arbitrary ways. Of course, humans are exposed to the same ethical conflicts, but we do have other resources available to us than our internalized 'rules'--including the power to communicate and negotiate challenging moral dilemmas with other moral agents.

Given the challenges of both top-down and bottom-up strategies for building safe, beneficial, and ethical AI systems, many have proposed the necessity of a 'hybrid' approach to machine ethics (Wallach and Allen 2009, Abney 2012, Arnold et al. 2017). Hybrid approaches incorporate both bottom-up machine learning of the complex patterns of moral behavior, and certain rule or principle-based ethical constraints upon such behavior.

One possible hybrid approach would seek to model machine ethics after the processes by which humans cultivate moral *virtues*, as described by Aristotle and other virtue ethicists. On this view, learning of ethical behavior begins by following a combination of simple rules ('don't kill,' 'don't lie, etc.') and habitually modeling the behavior of moral exemplars, but gradually produces a synthetic moral intelligence or "artificial practical wisdom" (Sullins 2016) grounded in an emergent capacity for increasingly adaptive and creative moral perception and reasoning. This prospect remains purely hypothetical, and may even require capacities for engineering synthetic emotion to address the 'frame problem' in machine ethics (Guarini and Bello 2012). Of course this may not be possible, desirable, or even ethical in its own right, if it requires creating artificial beings that can suffer.

The value alignment defender, however, may challenge the ambition of the hybrid approach as both *superfluous* (if the behaviorist IRL strategy can succeed at reliable value alignment), and potentially *dangerous* (since an agent with its own ethical intelligence has the freedom to modify its ethical thinking in ways that risk diverging from our morality and interests).

3. *Practical Differences*

The practical concerns *shared* by machine ethicists and value alignment proponents arise from the virtual certainty that AI systems will rapidly increase in number, scale, complexity, and power as they gain the capacity for increasingly autonomous operation. Already, AI systems are being developed for use in law enforcement, military, finance, education and healthcare contexts; mistakes in these domains will be costly not only in economic terms but in human suffering—and especially severe or cascading mistakes will endanger lives and institutions.

In this respect, the value alignment and machine ethics communities have a mutual mission to rapidly develop, refine, and scale new techniques for ensuring the safe, beneficial, and ethically responsible use of AI systems. However, their priorities begin to diverge when it comes to the contentious topic of artificial general intelligence (AGI) and/or superintelligence (SI), developments that a subset of AI researchers regard as inevitable (Kurzweil 2005; Applied AI 2018). Even if AGI/SI is only *modestly* likely to emerge in the future, many value alignment researchers believe that this prospect poses the kind of existential risk to humanity that makes even that small probability a cause for urgent preventative action. Just as we would not brush off a 1% chance of a massive asteroid striking the planet in this century, and would take immediate action to reduce that chance to zero, even a 1% chance of AGI/SI emergence is risky enough on this view to warrant immediate and intensive research efforts in AI safety and AGI containment strategies.

Many machine ethicists, in contrast, are far more skeptical about the prospects for AGI/SI--even in the long term--and regard the near-term ethical risks of narrow/task-driven AGI to be sufficiently acute *already* to deserve society's primary focus. That is, many machine ethicists worry that hype and fearmongering about AGI/SI and existential risk will distract policymakers and researchers' attention from the far more prosaic but also far more pressing AI risks that need to be addressed, and divert limited funding from efforts to manage those risks to far more speculative research programs of AGI/SI containment.

Machine Ethics and Value Alignment in Collaboration

Several themes of productive agreement emerged through the discussions at the Hastings Center workshops on *Control and Responsible Innovation in the Development of AI and Robotics*:

1. Value-alignment researchers and machine ethicists **share a deep commitment to AI safety for human beings and to ensuring that AI development remains compatible not only with long-term human survival, but with increasing human welfare and flourishing**. This top-level goal is more than sufficient common ground to warrant future collaboration and knowledge-sharing among these research communities, as well as respectful mutual critique and challenge to drive progress toward this goal. The *Control and Responsible Innovation* workshops at the Hastings Center laid new and solid foundations for this cooperation moving forward, bringing key representatives from each community into deep conversation and building greater understanding and trust among these researchers.

2. Because they share a common top-level goal, there is no intrinsic conflict between the two research agendas, and as long as there is sufficient research funding and talent to go around (admittedly a condition hard to meet in practice), progress among them need not be a zero-sum game. The value alignment and machine ethics research agendas can thus proceed independently as they have in the past, but since both have the potential to be informative to the other, **robust scientific collaborations could accelerate advances toward their shared goal**. For example, advances in techniques such as IRL for value alignment could be integrated in the hybrid systems advocated by some machine ethicists to ameliorate the brittleness of many machine moral logics. In turn, machine ethicists with expertise in ethical normativity could provide useful assistance to value alignment researchers who are seeking to make their reward and utility functions more nuanced, contextually sensitive, and appropriately responsive to the normative structure of human social behavior.

3. Both research agendas serve useful scientific purposes, namely developing and refining techniques for more complex forms of AI cognition and perception. Both serve a useful *social* purpose as well, by fostering more awareness of the limitations and risks of AI that need to be addressed before AI systems are entrusted with significantly more autonomy 'in the wild' than is safe and appropriate today. On this latter point, researchers from both communities should work together to **present a united front calling for social caution among policymakers and industry representatives** who may otherwise push for premature and unsafe levels of automation of safety-critical systems, or other systems that deliver key public goods.

The future of AI must be safe, beneficial to humans, and developed with social and scientific responsibility. If the conversations begun at the Hastings Center *Control and Responsible Innovation* workshops can continue in a spirit of collaboration, then value alignment and machine ethics proponents can help one another ensure that we reach that ultimate goal.

References

Allen, Colin, Iva Smit, and Wendell Wallach. 2005. "Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches." *Ethics and Information Technology* 7(3): 149-155.

Anderson, Michael, and Susan Leigh Anderson. 2015. "Case-Supported Principle-Based Behavior Paradigm." In *A Construction Manual For Robots' Ethical Systems: Requirements, Methods, Implementations*, edited by Robert Trappl, 155-168. Cham: Springer.

Applied AI. 2018. "373 Experts Opinion: AGI / Singularity by 2060 (2018 update)." February 15, 2018. Accessed July 18, 2018 at https://blog.appliedai.com/artificial-general-intelligence-singularity-timing/.

Arnold, Thomas, Daniel Kasenberg, and Matthias Scheutz. 2017. "Value Alignment or Misalignment? What Will Keep Systems Accountable?" *Proceedings of AAAI Workshop: AI, Ethics, and Society*. Accessed April 7, 2018 at https://hrilab.tufts.edu/publications/aaai17-alignment.pdf

Bringsjord, Selmer, Naveen Sundar Govindarajulu, Bertram Malle, and Matthias Scheutz. 2018. "Contextual Deontic Cognitive Event Calculi for Ethically Correct Robots (abstract)." Version 1025172359CA for ISAIM 2018. Accessed April 7, 2018 at https://www.semanticscholar.org/paper/Contextual-Deontic-Cognitive-Event-Calculi-for-Bringsjord-G./d519f2ae8c3a96709cca1c9e976519decf8bb836

Bostrom, Nick. 2009. "Ethical Issues in Advanced Artificial Intelligence." In *Science Fiction and Philosophy: From Time Travel to Superintelligence*, edited by Susan Schneider, 277-284. West Sussex: Wiley and Sons.

Govindarajulu, Naveen Sundar, and Selmer Bringsjord. 2015. "Ethical Regulation of Robots Must Be Embedded in Their Operating Systems." In *A Construction Manual For Robots' Ethical Systems: Requirements, Methods, Implementations*, edited by Robert Trappl, 85-99. Cham: Springer.

Guarini, Marcello and Paul Bello. 2012. "Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters." In *Robot Ethics*, Edited by George Bekey, Keith Abney, and Patrick Lin, 129-144. Cambridge, MA: MIT Press.

Kurzweil, Ray. 2005. *The Singularity is Near: When Humans Transcend Biology*. New York: Penguin.

Moor, James. 2006. "The Nature, Importance, and Difficulty of Machine Ethics," *IEEE Intelligent Systems* 21(4), 18-21.

Omuhundro, Stephen M. 2008. “The Basic AI Drives.” In *Proceedings of the First AGI Conference*, Vol. 171: Frontiers in Artificial Intelligence and Applications, edited by P. Want, B. Goertzel, and S. Franklin, 483-492. Amsterdam: IOS Press.

Russell, Stuart, Daniel Dewey, and Max Tegmark. 2015. “Research Priorities for Robust and Beneficial Artificial Intelligence.” *AI Magazine* 36(4), 105-114.

Sullins, John P. 2016. “Artificial Phronesis and the Social Robot.” In *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016*, edited by Johanna Seibt, Marco Norskov, Soren Schack Andersen, 37-39. Amsterdam: IOS Press.

Taylor, Jessica, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. 2016. “Alignment for Advanced Machine Learning Systems.” Technical report, Machine Intelligence Research Institute (MIRI). Accessed April 7, 2018 at https://intelligence.org/2016/07/27/alignment-machine-learning/.

Wallach, Wendell, and Colin Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

Towards Provably Beneficial AI

Stuart Russell

Early Developments

It hardly needs mentioning that progress in AI has been quite rapid over the last few years. For example, benchmark performance has reached or exceeded human levels in three areas—visual object recognition, speech recognition, and machine translation—that have long been considered among the "holy grails" of AI research. In the space of 24 hours, a single program, AlphaZero, became by far the world's best player in three games (chess, Go, and Shogi) to which it had no prior exposure.

These developments have provoked some alarmist reporting in the media, invariably accompanied by pictures of Terminator robots. Predictions of *imminent* superhuman AI are almost certainly wrong: several conceptual breakthroughs are still required. On the other hand, massive investments in AI research—several hundred billion dollars over the next decade—portend further rapid advances. Predictions that superhuman AI is *impossible* are unwise and lack any technical foundation.[1]

For these reasons, it seems prudent to assume that superhuman AI will eventually be achieved. We must, therefore, ask the question, "What then?" The generic answer—one given by Alan Turing himself[2]—is that creating entities more intelligent than ourselves leads to a loss of human control. The primary failure mode arises from machines optimizing fixed objectives that are not well aligned with true human preferences. As Norbert Wiener put it:[3]

> *If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively … we had better be quite sure that the purpose put into the machine is the purpose which we really desire.*

As Wiener also noted, this wasn't a problem as long as machines were stupid and had only small, local effects. We could always reset the machine and try again. When a machine is more capable than its human designers and connected to the Internet, this option probably won't be available.

As a foretaste of this, a relatively simple AI algorithm—the adaptive reinforcement learning system that maximizes click-through in social media—has already wrought havoc by coercing

[1] Moreover, similar statements of impossibility in other comparably important fields have turned out to be false. For example, many physicists in the 1920s and early 1930s pooh-poohed the idea of "atomic bombs." On September 11, 1933, Lord Rutherford described the possibility of extracting energy from atoms as "moonshine." On September 12, 1933, Leo Szilard invented the neutron-induced nuclear chain reaction.

[2] See, for example, Alan Turing (1951). Can digital machines think? Lecture broadcast on BBC Third Programme: "If a machine can think, it might think more intelligently than we do, and then where should we be? Even if we could keep the machines in a subservient position, for instance by turning off the power at strategic moments, we should, as a species, feel greatly humbled."

[3] Norbert Wiener (1960). Some Moral and Technical Consequences of Automation. *Science*, 131, 1355-58.

the political views of hundreds of millions of people toward extreme positions with the sole purpose of making them more predictable clickers.

At the time of the first Hastings Center workshop in April 2016, this *value alignment* problem was already well known. My research group at Berkeley had developed the basic idea of cooperative inverse reinforcement learning (CIRL)[4] as an approach to creating value-aligned AI systems and had seed funding from DARPA and FLI for this work. The framework is intrinsically game-theoretic. The simplest version is as follows:

- The world contains a human and a machine.
- The human has preferences and acts (roughly) in accordance with them.
- The machine's objective is to optimize for those preferences.
- The machine is explicitly *uncertain* as to what they are.

Shortly after the first workshop, the Open Philanthropy Foundation provided a major gift to set up the Center for Human-Compatible Artificial Intelligence (CHAI) at Berkeley, with branches at Michigan and Cornell. The core ideas were disseminated through a short paper in *Scientific American*[5] and a TED talk[6] as well as keynote talks at the main AI conferences, AAAI and IJCAI.

Findings

Overall, the three Hastings workshops were very helpful in refining the basic notion of provably beneficial AI and better situating it relative to the social-science tradition of attempting to pin down the notions of human welfare and morally correct action.

Value Alignment and Ethical Theories

The discussions led to a better understanding of how we came to be in such a problematic situation with respect to advances occurring within the classical definition of AI. The diagnosis goes right back to the beginnings of AI, which borrowed from a tradition of identifying human intelligence with goal achievement that stretches back at least to Aristotle.

The two basic steps in setting up the field of AI went (very roughly) like this:

1. We identified a reasonable notion of intelligence in humans: ***Humans*** *are intelligent to the extent that* ***our*** *actions can be expected to achieve* ***our*** *objectives.*
2. We transferred this notion directly to machines: ***Machines*** *are intelligent to the extent that* ***their*** *actions can be expected to achieve* ***their*** *objectives.*

Because machines, unlike humans, have no objectives of their own, we gave them objectives to achieve. The same basic scheme—optimizing exogenously defined objectives—underlies classical economics (utility functions), control theory (cost functions), statistics (loss functions),

[4] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell (2017). Cooperative Inverse Reinforcement Learning. In *Advances in Neural Information Processing Systems 25*, MIT Press.

[5] Stuart Russell (2016). Should we fear supersmart robots?. In *Scientific American*, 314, 58-59.

[6] Stuart Russell (2017). Three principles for creating safer AI. TED talk, Vancouver, https://www.ted.com/talks/stuart_russell_3_principles_for_creating_safer_ai

management science (various), and operations research (reward functions). Although this scheme is widespread and extremely powerful, *we don't want machines that are intelligent in this sense*.

Looking again at the definition of machine intelligence (2, above): we have no reliable way to make sure that *their objectives* are the same as *our objectives*. Like King Midas, we will inevitably fail to put the correct purpose into the machine. So let's try this instead:

> 3. Machines are ***beneficial*** to the extent that ***their*** actions can be expected to achieve ***our*** objectives.

This is probably what we should have aimed for all along. The difficult part, of course, is that our objectives are in us, and not in them. This is the general problem to which cooperative inverse reinforcement learning provides the germ of a solution. Optimal machine strategies in CIRL games turn out to be *deferential* to humans; for example, machines are motivated to ask permission, to allow themselves to be switched off,[7] and to act cautiously when guidance is unclear. Moreover, humans in this framework are motivated to act *instructively*—to (try to) teach their preferences to machines. Most importantly, under certain assumptions a machine that executes an optimal strategy in such a game has positive expected value for the human.

The Hastings discussions included many references to three main traditions in ethical theory: consequentialism, deontological ethics, and virtue ethics. Consequentialism is the idea that choices should be judged according to expected consequences. Deontological and virtue ethics are, very roughly, concerned with the moral character of actions and individuals, respectively, quite apart from the consequences of choices.

The criteria for what constitutes a satisfactory ethical theory for the design of machines are different from those that apply in the case of decisions made by humans. For example, consider two individual humans, Alice and Bob, whose actions are identical. Let's say each of them gives $1 to Connie, a homeless person with two young children. Alice does it automatically, without thinking—her reflexes are charitable, but they are reflexes; Bob, on the other hand, is genuinely moved by Connie's plight, considers the possible ramifications and alternatives such as giving the money to the local shelter, and then makes his decision. It seems reasonable to argue, as a virtue ethicist might, that there is intrinsic value in Bob's empathic reaction and thoughtful response. The same argument is much more difficult to make if Alice and Bob are machines completely lacking in subjective experience. It makes little sense to build machines that are virtuous or that choose morally valid actions if the consequences are highly undesirable for humanity. Put another way, we build machines to bring about consequences, and we should prefer to build machines that bring about consequences that we prefer. The nature of the internal processing—the specific sequence of computations—is of no moral consequence if the actual consequences are identical.

Many, Real Humans

The basic model outlined above requires many elaborations. These fall under two headings: satisfying the preferences of *many humans* and understanding the preferences of *real humans*.

[7] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell (2017). The off-switch game. In *Proc. IJCAI-17*, Melbourne.

Both topics form part of the staple diet of the social sciences. The workshops were extremely useful in making apparent the importance of this connection.
Machines making decisions on behalf of multiple humans face issues studied in the philosophy of utilitarianism and the economics of welfare aggregation, such as Nozick's utility monsters[8] and Parfit's Repugnant Conclusion.[9] Negatively altruistic human preferences such as "sadism, envy, resentment, and malice"[10] also cause difficulties; should machines simply ignore them? We added a new puzzle of our own: generalizing Harsanyi's famous social aggregation theorem,[11] we showed that the preferences of humans with heterogeneous beliefs must be weighted *dynamically* according to how well their predictions turn out.[12] This enables more flexible and efficient social contracts but has unsettling consequences for moral philosophy.

When dealing with real rather than idealized (rational or quasi-rational) humans, machines will need to "invert" actual human behavior to learn the underlying preferences that drive it. For example, chess players make mistakes because of computational limitations; a machine observing such a mistake should not conclude that the player *prefers* to lose the game. Other natural targets include the well-documented "heuristics and biases" in human decisions; the fact that human decisions occur within a hierarchy of individual and joint intentional commitments rather than *ab initio* optimization at each instant; and the role of emotions both in influencing human decisions and revealing deep underlying preferences.

The complex nature of human cognition raises a further question: when, if at all, is it possible consistently to attribute preferences to a sometimes non-rational entity? For example, Daniel Kahneman[13] argues that we have an *experiencing* self and a *remembering* self who disagree about the desirability of any given experience. Which one should the machine serve?

Finally, it is essential to consider the *plasticity* of human preferences, which obviously evolve over time through maturation, experience, and social influences. The social-media click-through catastrophe shows how rapidly machines can modify human preferences. There is a need for new philosophical analyses of rational preference change and for methods that prevent machines from *satisfying* human preferences by *modifying* those preferences to fit the status quo. This, in turn, leads to questions about equilibria of preference evolution and the potential for self-installation of pro-social preferences.

Practical Realization

[8] Nozick, R. (1974). *Anarchy, State, and Utopia*. Basic Books. Notice that humans have an *incentive* to appear to be utility monsters in order to gain a greater share of the machine's assistance.

[9] Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

[10] Harsanyi, J. (1977). Morality and the theory of rational behavior. *Social Research*, 44, 623–56.

[11] Harsanyi, J. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63, 309-321.

[12] Andrew Critch, Nishant Desai, and Stuart Russell (2018). Negotiable Reinforcement Learning for Pareto Optimal Sequential Decision-Making. In *Proc. NIPS-18*, Montreal.

[13] Kahneman, D. (2015). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

While the CIRL framework was developed as a way to avoid a failure mode for AI, it also has the converse property that CIRL systems are intrinsically desirable from a practical point of view: they provide provable benefits to the user and automatically generate a style of interaction that is highly appropriate for AI systems and quite difficult to produce by other means. For this reason, it seems desirable to demonstrate the benefits of the approach through paradigmatic examples that can convince the broader AI community to adopt the approach and to develop standards based around it. Such examples can also provide the opportunity to explore the preference structures of real humans in real settings; to begin to identify the major characteristics of human cognition that need to be considered when learning preferences from human behavior; and to surface concrete instances of situations where AI systems need to trade off the preferences of multiple humans, comply with laws, and so on.

The third workshop therefore devoted a significant amount of time to discussing a potential project to develop a personal digital assistant (PDA) for daily life. We would need to trust the PDA to use our credit cards wisely, to screen our calls and emails, to claim expense reimbursements, and to manage our finances. Obviously, a poorly designed PDA can do a lot of damage. It would not start out knowing the user's individual preferences, which leads directly to several concrete problems. These include inferring preferences from interactions with the user and with other, similar user (so-called *population IRL*); understanding user "commands" as evidence about user preferences, perhaps drawing on Doyle and Wellman's "ceteris paribus" semantics for goals;[14] and trading off the possibility and cost of making a mistake against the cost of asking the user for guidance. Furthermore, the PDA will necessarily be interacting with human other than its user, so it will be necessary to work out how to constrain its behavior to avoid stealing money or giving away passwords and how to trade off the user's preferences with those of other people so that the PDA is not constantly pestering others to benefit the user.

Governance and culture

At present, regulation seems to make sense only for specific use cases of AI, such as self-driving cars and insurance claims processing. It seems premature to propose broad regulations because we simply do not know what those might be, beyond perhaps tightening up areas where there may be ambiguity about liability for the actions of AI systems and legal confusion about causal chains. Eventually, it may be feasible to specify software design templates (perhaps based on CIRL agents) to which various kinds of applications must conform in order to be sold or to connect to the Internet, just as applications have to pass a number of software tests before they can be sold on Apple's App Store^{TM} or Google Play^{TM}. For example, adaptive algorithms that select advertisements or news items to display would need to be designed to avoid the systematic preference-modification effects of current methods. It would make sense also to create professional codes of conduct around the idea of provably safe AI programs and to integrate the ideas into the curriculum for aspiring AI and machine learning practitioners.

Whereas we are accustomed to the idea that pharmaceutical companies have to show safety and (beneficial) efficacy through clinical trials before they can release a product to the general

[14] Wellman, M. and Doyle, J. (1991). Preferential semantics for goals. In *Proc. AAAI-91*.

public, a "bunch of dudes chugging Red Bull"[15] at a software company can unleash a product or an upgrade that affects literally billions of people with no third-party oversight whatsoever. Governance structures seem to be essential to ensure that digital products are safe and effective.

Unfortunately, criminal elements, terrorists, and rogue nations would try to circumvent any constraints on the design of AI systems. The danger is not just that the evil schemes might succeed; it is also that might fail by losing control over poorly designed AI systems—particularly ones with evil intentions and access to weapons. This create a very serious policing problem. Already, we are losing the battle against malware and cybercrime. (A recent report[16] estimates over two billion victims and an annual cost of around $600 billion.) Malware in the form of highly intelligent programs would be much harder to defeat. A good first step would be a successful, coordinated, international campaign against cybercrime, including expansion of the Budapest Convention on Cybercrime. This would form an organizational template for possible future efforts to prevent the emergence of uncontrolled AI programs. At the same time, it would help to create a broad cultural understanding that creating such programs, either deliberately or inadvertently, is in the long run a suicidal act comparable to creating pandemic organisms.

Discussions at the workshops ranged beyond governance into the realm of culture and its interaction with AI. For example, the roughly one hundred billion people who have lived on Earth have spent roughly one trillion person-years learning and teaching, in order that our civilization may continue. Up to now, our civilization's only possibility for continuation has been through re-creation in the minds of new generations. (Paper is fine as a method of transmission, but paper does nothing until the knowledge recorded thereon reaches the next person's mind.) That is now changing: increasingly, it is possible to place our knowledge into machines that, by themselves, can run our civilization for us.

Once the practical incentive to pass our civilization on to the next generation disappears, it will be very hard to reverse the process. One trillion years of cumulative learning would, in a real sense, be lost. We would become uncomprehending passengers in a cruise ship run by machines, exactly as envisaged in the film *Wall-E* or E. M. Forster's *The Machine Stops*. Over the longer term, therefore, it will be essential to consider value of human autonomy and how to ensure that machines contribute to greater autonomy rather than greater enfeeblement. This is both a cultural as well as a technical project. Even if well-designed machines say no—that is, even if they insist that humans retain control and responsibility for their own wellbeing—myopic and lazy humans may disagree.

[15] Tegmark, M. (2018). Interviewed in the film *Do You Trust This Computer?*.

[16] https://www.securitymagazine.com/articles/88710-cybercrime-cost-600-billion-and-targets-banks-first

Section V. Agile and Comprehensive Governance

Problems arising from the lack of effective oversight of the digital technology industry have become of major concern in both the U.S. and in Europe. Information technology is dominated by giant corporations, and those same corporation will also dominate in the age of artificial intelligence. According to a June 2017 report by PricewaterhouseCooper, AI will drive a 14% increase in Global GDP by 2030, roughly $15.7 trillion. What kind of governance should be put in place to oversee the growth of the AI economy and the ever-expanding power of tech giants?

Cognizant of the mismatch between existing forms of governmental oversight and the demands fast developing emerging technologies were placing upon society, Gary Marchant and Wendell Wallach began to develop a new more agile and comprehensive model for technology governance in 2013. They referred to this new approach as a “governance coordinating committee” and proposed pilot projects for AI and robots and for synthetic biology. These fields were young and relatively unencumbered by laws, regulations, and competing standards.

At the first Hastings Center AI workshop, a cursory overview of that model was presented to the participants, who then brainstormed its application to AI. By of the third workshop, in the spring of 2018, it had become clear that a pilot project in the U.S. or the EU would be insufficient, and Wallach proposed an International Congress for the Governance of AI (ICGAI) to be convened in either 2020 or 2021. During the ensuing discussion, the workshop participants concluded that this timetable was too slow, and they unanimously proposed that an International Congress be convened within a year if possible. As one workshop expert proclaimed, “This is truly a case where proceeding with too much care is the enemy of the good.”

From the UN to the World Economic Forum, lack of agility in international governance has been noted as a major issue. Emerging technologies afford an opportunity to experiment with new forms of more adaptive and comprehensive governance.

A planning meeting for the proposed International Congress for the Governance of AI was convened in New York City on September 26, 2018, during the 2018 gathering of the UN General Assembly. The host/partners of the meeting were UN Global Pulse, The World Technology Network, and BGI4AI (Building Global Infrastructure for the Governance of AI). BGI4AI is a project started by Anja Kaspersen (presently Director of Disarmament Affairs at the UN in Geneva) and Wendell Wallach. Anja Kaspersen participated in all three of the Hastings AI workshops. Other participants from the workshops on the BCI4AI advisory board include Gary Marchant, David Roscoe, Francesca Rossi, and Stuart Russell.

The September 26, 2018, meetings was attended by 70 leaders representing major corporations and significant organizations in the AI ecosystem and international governance. These leaders enthusiastically endorsed the convening of ICGAI for November 2019. Furthermore, they proposed that the Congress move toward the establishment of an international body to oversee the governance of AI.

The attached papers explain the proposed approach for the governance of AI.

Recommendation: A consortium of industry leaders, international governmental bodies and non-governmental institutions, national and regional (e.g., the EU) 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 turned to 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 NYC when the UN General Assembly was in session.

An Agile Ethical/Legal Model for the International and National Governance of AI and Robotics

Wendell Wallach and Gary Marchant

Abstract—The accelerating pace of emerging technologies such as AI has revealed a total mismatch between existing national and international governmental approaches and what is needed for effective ethical/legal oversight. To address this "pacing gap" the authors proposed governance coordinating committees (GCCs) in 2015 as a new more agile approach for the coordinated oversight of emerging technologies. In this paper, we quickly reintroduce the reasons why AI and robotics require more agile governance, and the potential role of the GCC model for meeting that need. Secondly, we flesh out the roles for government, industry, engineering, and ethics in this comprehensive approach to the oversight of AI/robotics mediated by a GCC. We also propose a series of new mechanisms for enforcing (directly or indirectly) "soft law" approaches for AI through coordinated institutional controls by insurers, journal publishers, grant funding agencies, professional associations, courts and governments. Furthermore, significant attention must be directed to engineering and ethics solutions, including: AI safety, imbuing systems with values and the ability to make ethical/legal decisions, adequate testing and compliance procedures, and technology review boards. In light of the transnational nature of AI concerns and risks, we argue for an international GCC with complementary national and regional bodies. In addition, we show how a GCC can support and reinforce the governance initiatives of organizations such as the IEEE and the Partnership in AI. Finally, we propose convening a Global Congress as a first stage in establishing comprehensive effective oversight of AI and robotics.

***Index Terms*—artificial intelligence, regulation, governance, coordination, soft law, enforcement**

I. INTRODUCTION: THE NEED FOR AGILE GOVERNANCE

The accelerating pace of emerging technologies such as AI, and the onset of a Fourth Industrial Revolution, has revealed a total mismatch with existing governmental approaches and what is needed for effective ethical/legal oversight [1]. These emerging technologies exceed the regulatory scope, capabilities and jurisdiction of any one agency or nation. For example, AI raises ethical, legal and social concerns in need of governance relating to military use, safety, cybersecurity, privacy, transparency, bias, unfair business practices, antitrust, human enhancement, criminal justice, impacts on personal, family and societal relationships, economic equality, technological unemployment, existential risk, and no doubt many others. These diverse issues span many different industries, regulatory authorities, non-governmental organizations, experts and other stakeholders. While these concerns raise distinct issues that often must be addressed in their own way, they are also connected in that they relate to the same underlying technologies and therefore necessitate a more holistic approach.

In addition to the complexity of emerging technologies such as AI, the pace at which they are being developed also presents a major obstacle to traditional government regulation. AI is

W. Wallach is a Senior Advisor to The Hastings Center, Garrison, NY 10524 USA and chairs Technology and Ethics studies at the Yale Interdisciplinary Center for Bioethics, New Haven, CT 06520-8293 USA (e-mail: wendell.wallach@yale.edu).

G. Marchant, is Regents Professor and Lincoln Professor of Emerging Technologies, Law & Ethics, Center for Law, Science & Innovation, Arizona State University, Phoenix, AZ 85004-4467 USA
(email:gary.marchant@asu.edu).

developing at an accelerating trajectory, surprising even many AI experts about its recent speed and impact [2]. At the same time, our traditional governmental institutions of legislation, regulation and judicial review are slowing down rather than speeding up, creating the “pacing problem” [3].

To address these governance challenges, the authors proposed governance coordinating committees (GCCs) in 2015 as a new more agile approach for the coordinated oversight of emerging technologies such as AI [4]. Among its tasks, a GCC would comprehensively monitor development, flag concerns and gaps in oversight needing further attention, suggest means for addressing those gaps drawing upon an array of governmental and non-governmental mechanisms, and act as a good-faith broker mediating between the concerns of the various stakeholders. We further proposed that pilot GCC projects be started for AI/robotics and also for synthetic biology. The selection of these fields for pilot projects was occasioned by the fact that AI/robotics and synthetic biology are relatively new fields of research, largely unencumbered by rules and regulations.

Much has happened since our initial proposal. Machine learning approaches have led to breakthroughs in AI, and CRISPR/Cas9 has speeded up gene editing and in turn the development of genomic products and synthetic organisms. Governments have taken notice and are studying ways to regulate AI and genomics, and a variety of governance proposals have been put forward both within and outside governments around the world. In this paper, we will focus upon those applicable to the development of AI.

The European Union (EU) and many individual countries have begun discussing laws and regulations, agencies, enforcement regimes, and other governmental mechanisms for the oversight of AI and robotics. The “Civil Law Rules on Robotics European Parliament resolution”[5] proposed by the EU Parliament which recommends the establishment of a European Agency for Robotics and Artificial Intelligence and the General Data Protection Regulation (GDPR) enacted by the EU are particularly noteworthy. The proposal for a robotic agency is robust and even incorporates some of the functions we propose for a GCC, but is more regulatory in its focus than the GCC. This may well be appropriate for the EU. Whether the EU will act on these recommendations and actually establish such an agency is unclear, as the EU Commission has put forward a more limited pathway in its recent Communication on AI [6]. Proposals for draft guidelines directed specifically at artificial intelligence are expected from the EU before the end of 2018. We certainly welcome experimentation with many forms of oversight, but we also question the applicability and desirability of a more regulatory-centric approach for many other nations.

Japan, South Korea, and Singapore have their own independent efforts, and other nations are quickly following suit. The United Arab Emirates has even appointed a Cabinet-level Minister of State for Artificial Intelligence. From Africa to Asia, nations and cities are formulating plans and building in 21st century infrastructure that will ensure their municipalities are Smart Cities [7].

In addition, many existing and new non-governmental organizations (NGOs or civil society), alliances, and research centers have sprung up to address potential benefits, societal impacts, risks and dangers posed by the deployment of AI. For example, the IEEE has begun formulating international standards. The United Nations Secretary General is convening a high-level panel on digital cooperation. The World Economic Forum has also initiated a high-level council on artificial intelligence. The Partnership in AI (PAI), a consortium of representatives from leading companies, non-governmental organizations and research centers, has begun underscoring best practices.

As governmental and non-governmental responses to advanced technologies emerge, our initial proposal for a GCC has also evolved. The GCC model was originally proposed as a new mechanism for the governance of emerging technologies in the U.S. In contrast to the proposal for a new government agency (or "commission") to coordinate oversight of robotics by Ryan Calo [8], we argued that governmental oversight is necessary but not sufficient for AI governance.

We felt institutions similar to a U.S. GCC would be needed in other nations and that some form of international coordination between these national bodies would be required. The increasing focus on international concerns gave birth to a project to Build Global Infrastructure for the Comprehensive and Agile Governance of AI (BGI4AI) in the spring of 2016. A year later, BGI4AI proposed convening an International Congress for the Agile and Comprehensive Governance of AI.

To date no country has instituted a GCC. Nevertheless, the model has been discussed and praised in a wide variety of forums including at the UN in Geneva and the World Economic Forum in Davos. Furthermore, it has been discussed as a means to address a wide variety of international challenges including food security, biosecurity, and the governance of geoengineering and of the oceans, to name a few. Whether an actual institution that embodies all the main features of the GCC model ever comes into being is less important than the fact that the ideas inherent in this model are facilitating reflections on creative means for the governance of a broad array of challenges.

In this paper, we expand on our initial proposal to describe how a GCC, perhaps at an international level, could help to break the current logjam with respect to agile and effective governance of AI. We conclude with a reiteration of the proposal to convene an International Congress to establish mechanisms for the agile and comprehensive governance of AI.

II. The GCC Approach

The basic idea behind the GCC is that an orchestra needs an orchestra conductor – not to play the instruments for the various players, but to coordinate all these important parts of the performance. For emerging technologies like AI, there is an explosion of governance strategies, actions, proposals and institutions. All have an important role to play – whether they are from government, industry, NGOs, academia and research centers or some combination of the above, but no one entity or program can hope to govern the fields of AI and robotics *in toto*. What is missing is some mechanism for communication, coordination, synchronization and synergy.

It is for this reason we proposed the idea of an "issue manager" for the governance of individual emerging technologies like AI, which we named a Governance Coordinating Committee or GCC [4]. The GCC would serve several coordinating functions. One function would be as an information clearinghouse, by collecting and reporting in one place all significant programs, proposals, ideas or initiatives for governing AI. The GCC could also perform a monitoring and analysis function, such as identifying gaps, overlaps, and inconsistencies with respect to existing and proposed governance programs. It could serve as an early warning system, by noting emerging problems that are not addressed or covered by existing governance programs. It could provide an evaluation program that scores various governance programs and efforts for their metrics and compliance with stated goals. The GCC could provide a forum for stakeholders to meet and discuss governance ideas and problems and to produce recommendations, reports, and roadmaps. It could serve as a trusted "go-to" source for the media, the public, scholars and stakeholders to obtain information about AI and its governance. Finally, the GCC could serve as a convener for interested stakeholders on specific problems to meet and try to forge a negotiated partnership program for tackling unaddressed problems or governance needs.

There are many practical implementation challenges that need to be addressed for the creation of a GCC. Who would fund it? What type of governance system would be needed to operate the GCC? How would the GCC be evaluated, and by whom? How much staff would the GCC have, and who would hire them? Would the GCC have a direct or indirect government role, or would it be solely outside of government? How would stakeholders have a say and role in the operation of the GCC? What would be the specific goals and functions of the GCC? These are critical questions, but they do not lend themselves to one obvious set of answers. Rather, they are challenges that need to be negotiated and discussed in the context of a specific proposal and effort to create a GCC, and by as broad a range of stakeholders as possible.

To facilitate this creative process, we extend and elaborate on our original GCC proposal here to provide some additional insights, timely possibilities and benefits that a GCC could play in the governance of AI.

III. Enforcement

In our initial GCC proposal we emphasized the importance of soft governance mechanisms, which include industry standards, professional society codes of conduct, laboratory practices and procedures, insurance policies, statements of principles, voluntary government programs, certification programs and similar measures. Soft law measures impose substantive expectations or obligations that are not directly enforceable by government. We suggested that soft governance mechanisms should be favored in the GCC over hard governance (laws, regulations, regulatory bodies and courts) because they often involve multi-stakeholder participation, and can be adopted and modified more quickly and nimbly than traditional legal instruments. Another benefit of soft law mechanisms is that because they are usually not associated with a specific regulatory agency or jurisdiction, they can be applied at the international level [9].

However, the obvious weakness of soft governance mechanism lies in the difficulty, if not inability, to enforce them directly. There are nonetheless a number of indirect ways to enforce soft law measures, and the GCC can provide an appropriate forum for bringing the relevant players together to implement such soft law enforcement mechanisms. For example, we propose an additional and new role for governments, which is to create means to punish those who violate soft governance standards in a manner that leads to harm to people, non-human animals, the environment and institutions, or establishes practices that have undesirable societal impacts. Such an indirect government enforcement opportunity can be created using soft law instruments, perhaps negotiated or ratified through the processes of the GCC.

While the specific legal authority for such a government role will vary country by country, an example is provided by the Federal Trade Commission (FTC) in the United States. The FTC has a long-standing statutory authority to take enforcement action against "deceptive and unfair" business practices. The FTC has in recent decades re-interpreted this authority to apply it to companies that fail to comply with their publicly stated commitments, including adherence to soft law instruments such as privacy standards or codes of conduct [10]. The FTC's legal position is that a company's failure to live up to its public commitment misleads and deceives consumers, in violation of the statutory prohibition of deceptive and unfair practices. A GCC could help create or promote a private code of best practices for AI and robotics that participating companies would agree to, with the understanding that the FTC, for example, is empowered to take enforcement action against companies that fail to comply with their commitments. Similar indirect governmental enforcement mechanisms may be possible in other jurisdictions, or through similar agencies in other countries.

Courts may also have some enforcement role for soft law instruments created or publicized by GCCs. Private standards can set the standard of care for industry actors, particularly in the absence of any regulatory standards. Thus, private standards can provide a partial liability shield for those entities that comply with the standards, and can be used as a sword to establish the lack of due care by those entities that fail to comply with the private standards [11]. The more recognized and accepted the private standard, the more force it has as a shield or sword for liability in personal injury or other tort lawsuits. GCC endorsement of a soft law instrument could therefore give it more salience in private lawsuits, and could provide another indirect enforcement mechanism.

In addition, many governments and major corporations strengthen standards by requiring that they be met for products and services purchased by the government and industry. ISO 9000, for example, is an international quality management and quality assurance standard designed to increase business efficiency and the quality of products. Organizations that demonstrate the ability to provide products and services consistently can apply for ISO 9001 certification (a subset of ISO 9000). While organizations that fail to meet ISO 9000 standards are not directly punished, they are indirectly punished in their inability to sell their products and services into large markets.

Insurance companies also provide an indirect enforcement mechanism. After the asbestos debacle [12], liability insurers realized they cannot afford to insure companies or products that present unknown and potentially unlimited liability if harms occur. As a result, liability insurers are increasingly taking a more active risk management role for emerging technologies that present highly unknown but potentially widespread risks. For example, liability insurers for companies that manufacture or handle nanotechnology materials are increasingly requiring their clients to adopt an active risk management program as a condition for coverage [11]. These risk management programs often involve a commitment to comply with a voluntary standard or code of conduct. A GCC could work with companies and insurers to identify an appropriate set of risk management standards for companies working with AI applications that present significant risks.

There are other mechanisms for indirectly enforcing soft law instruments that a GCC could help facilitate. Journal publishers could agree to only publish articles that comply with applicable codes of conduct or professional standards. Funding agencies could condition funding on compliance with appropriate soft law standards. Research institutions could mandate compliance with soft law standards by their employees, perhaps enforced by an institutional review committee based on Institutional Biosafety Committees [13]. All of these mechanisms hold significant potential for indirectly enforcing soft law norms, and a GCC could provide the impetus and focus to enable such efforts.

IV. Top-down and Bottom-up Governance Mechanisms

The agile and comprehensive governance of AI and robotics will need to encompass both top-down and bottom-up approaches to ensuring that systems are safe and act appropriately. By top-down mechanisms we are referring to hard law, regulations and regulatory institutions as well as soft governance mechanisms. Bottom-up refers to the processes through which AI systems are developed and deployed. We refer to these as process soft governance and include both engineering solutions and means by which a company engages is self-governance.

A.) Hard and Soft Law

GCCs (international, regional or national) provide both top-down and comprehensive horizontal mechanisms for the governance of AI and robotics. Whether the GCC is within an existing governmental body, such as the UN or the European Union, or is a non-governmental institution, it will function as a largely horizontal oversight body monitoring developments comprehensively and facilitating a loose coordination of the various institutions addressing specific concerns. Most AI and robotic applications are context specific and are likely to be governed by vertical industry-specific institutions. For example, a rich body of laws, norms and institutions already exist for ensuring that healthcare is in the best interest of patients and that public health concerns, such as the likelihood of a pandemic, are managed effectively. Most applications of AI and robotics within healthcare will fall within preexisting bioethical and legal standards. The few new challenges these fields pose for healthcare are likely be taken up by the healthcare industry. The widespread and fair dissemination of healthcare benefits from AI will hopefully also be attended to by existing international institutions, e.g., World Health Organization and the International Committee of the Red Cross, and philanthropic organizations, e.g., the Gates Foundation. But even in AI's impact on healthcare there may be some need for coordination when, for example, guidelines or court rulings conflict.

In addition to its coordination function, the role of the GCC will be more toward monitoring gaps and ensuring they get addressed, and in facilitating work on broad societal challenges such as international security and the weaponization of AI, existential risks, technological unemployment, and normal accidents (system risks) that can cascade and potentially destabilize political and economic institutions. Whether fulfilling these functions will require GCCs to have broad legislative and executive powers may depend upon the responsibilities taken on by other institutions. The IEEE, for example, in its widely disseminated *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* goes beyond the setting of standards to offering a broad array of policy recommendations. Konstantinos Karachalios, the Managing Director of IEEE-Standards Association, foresees the prospect that IEEE takes on some governance responsibilities for AI, but whether it will actually do so is unclear at this time.

Top-down governance works best when complemented by bottom-up oversight, and here to there is need for new mechanisms, in particular, technology review boards (TRB) and corporate AI officers. Central to the GCC model is the importance of first evaluating engineering and ethical solutions to challenges before turning to either soft law or hard regulations. Indeed, a broad array of challenges can be met through value added design, imbuing AI systems with capabilities for making ethical and legal decisions, safety and control features, the reinforcement of existing societal values, and the establishment of new norms. Nevertheless, appropriate procedures must be developed to evaluate the feasibility of addressing gaps through technical solutions, ethics, social engineering, or corporate responsibility.

Hard and soft law can enforce the need for such procedures, but they are most effective when there is also an impetus for them to evolve from the bottom up. The good news is that, unlike past generations, many AI and robotic researchers seriously embrace the need to be socially responsible scientists. Furthermore, industry leaders recognize that AI, similar to genetically modified foods, could potentially be rejected by a large segment of the public. The rejection of genetically modified foods has been codified in jurisdictions such as Europe and many African nations. AI researchers and industry leaders are also particularly sensitive to automations impact on jobs and wages, as well as futurist fears about existential risks posed by artificial general intelligence (superintelligence) [14]. These initial concerns, however, have spread to embrace nearer-term problems including AI safety, data security, biased algorithms, and responsible innovation.

The rewards, and therefore the pressure, of being the first to innovate and deploy are real. Therefore, there will also be a temptation to cut corners and deploy systems that have only been tested in a cursory manner. Compliance requirements and technology review boards are one means of cutting down on the deployment of risky systems.

Some management teams may be satisfied with the appearance of being responsible. But responsible innovation is not about public relations. It relies on concrete practices, that have been created, implemented, tested and refined by industry leaders who possess vision and foresight.

B.) Process Soft Governance

In the field of artificial intelligence there are many risks, societal impacts, and ethical concerns that can be addressed in the way a system is engineered and through corporate review of technologies under development.

1.) Engineering and Ethics

In the imagination of some engineers, all problems can be solved technologically. But often their proposals are based upon fanciful gadgetry that is not feasible with the available tools and techniques. Determination of the feasibility, timeliness, and cost of developing technological solutions can be weighed against other options such as social engineering or government regulation. The work necessary to make these determinations is likely to be done by scholars or engineers, but if it has not been performed, a GCC could flag the lack of adequate analysis.

The education of engineers and their commitment to follow best practices and compliance standards is primary. Their full understanding of engineering ethics can ensure that the systems they design will inculcate prevailing norms and have a positive societal impact. This value-added design process can also be facilitated by integrating ethicists and social theorists into design teams, not as naysayers, but as fellow designers sensitive to ethical and societal concerns. For example, systems can be designed to maximize privacy. Determinations at the outset as to who should be held responsible if a system fails could direct engineers to change the platform upon which the system is built.

The prospect for developing AI systems and robots capable of factoring norms, ethical concerns and laws into their choices and actions has received significant attention by science fiction writers, then philosophers and computer scientists [15], and has now become an engineering challenge referred to as machine morality, machine ethics [16], or value alignment [17]. This is an exciting research trajectory. However, progress will be slow. Moral decision-making machines will initially be designed for bounded applications in which the options they confront are limited.

2) Technology Review Boards and Corporate Oversight

In support of standards there will be need for mechanisms to ensure compliance, testing, and the overall safety of systems before they are deployed. These will depend upon the commitment of corporations, universities, and other institutions to responsible research and innovation. The threat of hard governance mandates can be used to elicit industry-backed initiatives and a commitment to effective soft law. Such industry-backed initiatives should include testing procedures and certification for systems in compliance with safety standards set by international bodies such as the IEEE and ISO.

Safety has many dimensions from engineering specifications that lower the possibility of fires, electrical shocks, or other harms; to inspection regimes and assurance policies that certify the maintenance of systems; and to ethical standards for the treatment of research subjects. Research and ethics compliance standards already exist for research with human subjects and non-human animals or research utilizing hazardous materials or posing environmental risks. A few countries back up research compliance standards with review boards, which must be satisfied that designated ethical procedures and safeguards are in place before the research can proceed. These are sometimes seen as slowing down research and overly-bureaucratic. But countries throughout the world, even those without the resources to put in place rigorous oversight mechanisms, wish to protect their citizens from being used unethically as subjects by unscrupulous researchers. Foundations for human subject research were set by the Nuremberg Code established after World War II and have been built upon to meet the needs of individual countries. The Office of Human Research Protection within the U.S. Department of Health and Human Services publishes reports that compile international human research standards. The 2017 compilation covered 126 countries and enumerated more than 1,000 laws, regulations and guidelines [18].

If research in AI or robotics involves human or non-human animal subjects it should be subject to review and the pre-existing codes. However, for a variety of reasons this research often escapes comprehensive ethical review and systems are deployed without full consideration for their impact on humans, non-human animals, the environment, and institutions. For example, a robot developed to care for the homebound and elderly may never be exposed to actual subjects during its development. A machine learning system might be trained on an available database, but its output will not have any impact on humans until it is actually deployed.

Once deployed an AI application or a robot can have a dramatic societal impact that often goes unnoticed until after it is deeply embedded in the fabric of daily life. Consider, for example, the use by third parties of social media to influence elections around the world.

If industry leaders are serious about the responsible development of AI and robotics, they should establish technology review boards (TRBs) that go beyond research ethics and assess the impact the tools and techniques they are developing will have once deployed. Among the activities a TRB would engage in is the consideration of worse case scenarios, disaster planning, determination of who might be held responsible when a system fails, the fairness and privacy implications of the data that the system will use, and analysis of societal impacts should the system be widely disseminated. On the one hand, a TRB could assess the liability of a company for systems it markets and protection from class-action lawsuits. On the other hand, a TRB will ensure that the technology is developed responsibly. Reports from review boards get easily lost or are ignored, so the TRB should include a corporate AI Ethics Officer (AIEO), with the power to bring concerns to the attention of management and the board of directors.

In addition, we recommend a cross-industry consortium that sets standards for testing systems and certifies those that meet the highest standards.

There are many reasons why a TRB for AI and robotics is needed. Some of these reasons apply to other technologies and earlier innovations. These reasons include:

a) Overlooked Impact – Unless it is explicitly considered in advance, the influence of AI systems on citizens and on society can easily be overlooked. For example, it has recently been recognized that output of AI systems can be biased if the system has been trained on data that contained implicit biases. An AI system approving loans demonstrated prejudices even when the race of applicants was not noted in their personal record. The problem of algorithmic bias has now been underscored and yet those testing and deploying machine learning systems may not notice a particular bias unless they actually test for it before deploying the AI application.

b) Testing Limitations: Testing is expensive, and testing for all possible contexts in which an AI application might be deployed can be time-consuming and extremely costly. Few companies have adequate time and money to test systems thoroughly. Often, they make basic judgements as to what to test for, deploy systems, and wait for feedback from users about problems. This, in effect, makes the users subjects of research. Unfortunately, the ethical requirements for informed consent for research subjects are not extended to users once the system is deployed.

c) Machine Learning: Each new strategy or task learned can alter a system's behavior. Furthermore, learning can alter the very algorithm that processes information. Deploying a system that can change its programming significantly for a potentially harmful or mission critical application is risky. Furthermore, few enterprises have the resources to constantly retest systems that learn.

d) Unpredictability: Robots and AI systems are best understood as complex adaptive systems that are deployed in complex socio-technical contexts. Complex adaptive systems are subject to occasional unpredictable behavior, and periodically such unanticipated behavior can be harmful.

e) Low probability events: Low probability events do occur, and yet their likelihood is often minimized or dismissed. Integrating features into systems that minimize the damage from low probability events adds cost, but this expense can be insignificant in comparison to the cost of a potential disaster.

f) Normal accidents: Charles Perrow coined the terms "normal accidents" or "system accidents" to capture the likelihood of unanticipated events from complex tightly coupled systems even when no one does anything wrong. The near-nuclear meltdown at Three Mile Island in 1979 is an example of a normal accident. Tightly coupled elements of a system can cascade into a series of failures. AI systems will be highly prone to creating system accidents. By calling such accidents "normal" Perrow creates a contrast to the false notion that such systems can be highly reliable [19].

g) Moral machines: AI systems capable of factoring legal and ethical considerations into their choices and actions will be particularly difficult to test and certify. Initially such systems will function within very constrained contexts, yet as their autonomy expands to other realms, the likelihood, of unanticipated behavior can expand exponentially.

Not every contingency can be anticipated. Nonetheless, anticipatory planning can lower the likelihood of harmful behavior and undesirable societal impacts. This in turn protects corporate interests. For all of these reasons we recommend that the AIEO and TRB continue monitoring systems after they have been deployed as a precautionary measure, and to catch unintended risky behavior.

Whether a governance coordinating committee provides recommendations and oversight for TRB and AI ethics offices will depend largely upon whether or not these functions are taken on by other bodies. Paul Daugherty, the CTO of the corporate consulting firm Accenture, for example, has been recommending that corporations create AI ethics officers. Helping corporations put in place effective AI oversight is clearly a service Accenture hopes to market, and in turn they and other consulting firms are likely to develop a catalog of best practices for those companies dedicated to the responsible development and utilization of AI tools and techniques.

V. International and National Oversight of AI and Robotics

In the spring of 2016, a project to Build Global Infrastructure to Ensure that AI and Robotics are Beneficial was initiated [20]. This project is referred to as the BGI project or BGI4AI (https://bgi4ai.org/). The BGI project is a pilot for applying the GCC model to the ethical/legal oversight of these two fields of research (AI/robotics), but while a GCC was initially proposed for the U.S., this new project begins as an international program with complementary national or regional bodies. A complementary GCC could cover the needs of one country or a region, such as a Pan-Arab governance coordinating committee. For theoretical purposes the international body might be referred to as a global governance coordinating committee (GGCC). But once initiated, whether within the United Nations, the IEEE or as a new NGO, we expect the monitoring, multi-stakeholder engagement, coordinating, and other functions to be established under a new name. A central role of a GGCC and its complementary national or regional GCCs will be to underscore gaps in existing mechanisms for the oversight of AI/robotics, to propose new mechanisms to address those gaps after considering and evaluating an array of available tools, and to build the governance infrastructure necessary to sustain and perhaps even enforce those proposals.

Some of the concerns AI and robotics pose must be addressed globally while others are better left to regional, national or local ethical/legal oversight. For example, lethal autonomous weapons, and whether their deployment should be restricted by an arms control treaty, can only be resolved internationally. The Convention on Certain Conventional Weapons at the United Nations (UN) in Geneva has already taken up this question. Regulatory policies for the deployment of fully autonomous vehicles are being developed by many countries independently, and even states or regions within those countries.

While each country could in theory establish its own technical standards, testing procedures, compliance requirements, and quality management standards that must be met before products can be marketed, commonly they adopt those developed by international standard-setting bodies such as the IEEE and ISO. For other regulatory and soft law concerns, many countries are unable to establish their own requirements, or adopt those set by other countries. One role for a GGCC might be to underscore "best practices" and outline considerations for various national and regional bodies as they consider the most appropriate soft and hard law for their culture. Indeed, these "best practices" might even be considered *de facto* international standards, subject to variations introduced by national and regional GCCs. This would be particularly helpful for poorer regions. For example, international standards for autonomous vehicles could facilitate their deployment and ensure even nations without their own regulatory policies are protected.

The BGI project will begin with the establishment of a GGCC and several independent yet complementary GCCs.

VI. Support and Coordination, Not Competition

AI is beginning to affect every facet of modern life, and as it does so, an array of existing institutions and a proliferation of new centers and consortia have arisen to tackle emerging challenges. How can a GGCC or GCCs support and comple-ment (not compete with) the many international NGOs (e.g., the IEEE), governments or economic and political unions (e.g., the EU), industry promoted consortiums (e.g., the Partnership in AI), and research centers (e.g., AI Now) that have emerged to address challenges arising in the development and deployment of AI and robotics? Ethical guidelines, standards, principles, protocols, policy recommendations, research findings, tools for data analytics, and technical means to ensure safety and fairness are appearing, and will continue to be developed by these various initiatives.

Most of the initiatives are siloed attempts to deal with specific concerns or research centers whose influence will be limited unless its recommendations come to the attention of policy makers and industry leaders. None of these entities or programs can hope to govern the fields of AI and robotics *in toto.*

Nevertheless, each body is sensitive to competition from similar institutions, and will be unwilling to participate in joint deliberations if it feels the deliberating body will merely usurp its ideas and authority. In order to be effective a GCC (or GGCC) must attract the involvement of these other institutions, respect and support their contributions, and provide services that they cannot provide by themselves. It should not usurp the authority of other institutions. Rather, it should support their activities and facilitate their working together to ensure that best practices come to the fore and that the resources of individual institutions are not wasted through unnecessary duplication of effort. It will be helpful if the various institutions are aware of similar projects being performed by other researchers and institutions. Furthermore, it will be helpful for those proposing new policies, standards, and guidelines to be able to bring their work to the attention of others who have a good prospect of effecting their adoption.

In addition to governments, a few of these bodies have, or are expected to have, significant worldwide impact on the development of AI and robotics. Among the most influential internationally are the IEEE and the World Economic Forum. The Partnership on AI (PAI) – formed by Amazon, Apple, Facebook, Google, IBM, and Microsoft – is young, and yet it is casting a wide net and has already embraced many NGOs and research laboratories. PAI will certainly be influential, but it remains unclear whether this initiative will eventually include representation from all regions and all industry leaders. A few programs have also begun within the UN to address concerns posed by AI. But, as of this writing, no one institution can claim to speak for or include all the key stakeholders and on the broad array of issues arising from the development of AI and robotics.

Furthermore, much of the focus on the various emerging concerns is dominated by industry leaders and researchers from Europe and North American, as well as Japan and South Korea. Meanwhile, the BRICS economies (Brazil, Russia, India, China, and South Africa) have to date been less active in international forums on AI and robotics. Furthermore, China and it primary IT corporations (Alibaba and Baidu) rival counterparts in the U.S. in the development of AI. The Arab world, Africa, South and Central America have still to make their voices fully heard. In other words, there is a need for a GGCC.

A GGCC must draw upon and serve all of the stakeholders from industry and civil society to governments and international standard setting bodies. Hopefully, it will also find means to represent the interests of under-served nations and people, even those that lack stable governments. There is no shortage of opportunities for the fruits of AI and robotics to benefit all of humanity, but this can only occur when risks, dangers, and undesirable societal impacts are also being mitigated. A comparable level of responsibility will fall upon national and regional GCCs.

VII. Outcomes Not Merely Process

The GCC model offers a process and a framework for responsive and agile governance. The details of putting a GCC or GGCC into place are extensive and will be complicated, given the fact that AI and robotic applications are context specific, and each context, such as healthcare, will require its own supporting mechanisms and institutions. An agile process, however, is only worth pursuing if it effectively leads to significant outcomes. The development of the supporting institutions must proceed hand-in-hand with the pursuit of specific goals. Furthermore, the goals will help dictate the structure of the institutions and mechanisms put in place. The challenge lies in forging mechanisms that will serve both immediate goals and longer-term needs for responsive and agile oversight. With this in mind we will propose a project for the creation of a GGCC and the first national and regional GCCs.

Three near-term issues have emerged regarding the fairness, transparency, and integrity of machine learning systems [21]:

a) There is a lack of transparency/opacity as to how neural networks achieve their outputs, i.e., reach conclusions. This is particularly problematic given the explosion in use of deep-learning algorithms for a vast array of applications. Should an accident or harms occur, there may be no way to even forensically determine what went wrong.

b) The output of a deep learning algorithm will be unfair and biased if there is bias inherent in the dataset the system is trained upon. A machine learning algorithm might also yield a false or dangerous output if the data it is fed is filled with inaccuracies or simply too limited (insufficient in depth) to reach an accurate conclusion [22].

c) Concentration of power, claims of data ownership, and the inability of individuals to opt out of IT services whose handling of their data can jeopardize their security, privacy, and finances is receiving considerable attention in the media and by some governments. These issues are exacerbated when the use of an individual's data by social media companies makes users susceptible to political and ideological campaigns, behavior manipulation, and undesired marketing campaigns.

AI engineers, data analysts, standard-setting institutions, multi-stakeholder forums, scholars from a variety of disciplines, research centers, policy planners and some governments are working on means to address various facets of these challenges to try to mitigate potential harms. A General Data Protection Regulation, which comes into force in 2018, has been enacted by the EU. Among other provisions, these regulations provide for a right to obtain an explanation for decisions made by an algorithm, as well as the right to opt-out of various forms of data collection. Arguably the EU's regulations on these matters may be too broad and may even unnecessarily stultify innovation and economic progress. Nevertheless, they underscore the importance attributed to such issues.

Moreover, the landscape is changing. Policy makers will hopefully clarify when a lack of algorithmic transparency is problematic and when it is not. Data analysts are likely to produce tools that help illuminate biases and other limitations inherent in training data, as well as biases in system outputs. AI researchers are developing technical tools for illuminating the processes whereby "black box" systems arrive at their output. These promise to provide a degree of transparency, forensic capabilities, and some capacity to explain the reasons for system conclusions. The difficulty lies in the fact that while research advances and standards are being formulated, leading industry players, healthcare providers, legal decision-makers, and other parties are rapidly deploying systems and marketing products whose safety and societal impacts have not been determined.

We propose an International Congress for the Governance of AI (ICGAI) as a forerunner to the establishment of a GGCC. Focusing on the issues of algorithmic transparency and algorithmic bias are particularly appropriate as agendas for such an international gathering. For example, this Global Congress would establish preliminary guidelines for the deployment of algorithms that are not fully transparent. It would clarify when learning systems can be exempt from transparency requirements, what testing and compliance must be performed before potentially risky systems are deployed, and in which situations or contexts systems that lack transparency (opacity) should never be deployed. The standards, practices and procedures for setting these preliminary guidelines may have already been developed by other bodies. However, these other players are likely to be dominated by industries and institutions concentrated in North America or Europe. It therefore becomes important for additional companies, institutions, countries and regions to evaluate whether such guidelines are appropriate given their needs. In other words, the Congress provides an opportunity for stakeholders to endorse (or modify if necessary) the best practices that have emerged to date.

We propose that this ICGAI be held in a nation and at a venue that is considered relatively neutral. Given the dynamic state of the research and lessons learned from monitoring best practices, preliminary guidelines will need to be revisited in a few years, modified and hopefully made more precise. The very act of convening an International Congress provides an opportunity to lay foundations for multi-stakeholder oversight of AI and robotics. The ICGAI itself will hopefully endorse steps towards building agile and responsible institutions for the continuing oversight of AI and robotics. Whether this continuing oversight takes the form of establishing a new non-governmental organization similar to a GGCC is less important than that the coordinated monitoring of research and system deployment of AI will have begun.

As the first step toward the proposed ICGAI, an initial planning meeting was convened in New York on September 26, 2018, while the UN General Assembly was in session. The meeting was co-hosted by UN Global Pulse, the World Technology Network (WTN), Building Global Infrastructure for AI (BGI4AI), and NYU Stern School's Digital Governance Initiative. It was attended by over seventy of the leaders and/or high-level representatives of the key AI-related and governance-focused organizations in the world. The attendees enthusiastically endorsed moving forward on planning an ICGAI for late in 2019.

References

[1] G.E. Marchant, and W. Wallach, "Introduction," in *Emerging Technologies: Ethics, Law and Governance,* 1-12. London, U.K.: Routledge, Nov. 4, 2016.

[2] Executive Office of the President. Artificial Intelligence, Automation, and the Economy, December 20, 2016. [Online]. Available: https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF

[3] G.E. Marchant, "The Growing Gap between Emerging Technologies and the Law," in *The Growing Gap between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem* Dordrecht: Springer, 2011, pp. 19-33.

[4] G. E. Marchant and W. Wallach, "Coordinating Technology Governance," *Issues in Science & Technology*, 31[4] pp. 43-50, Summer 2015.

[5] European Parliament Resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)). Available: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2017-0051+0+DOC+XML+V0//EN

[6] Communication From The Commission To The European Parliament, The European Council, The Council, The European Economic And Social Committee And The Committee Of The Regions: Artificial Intelligence for Europe, Apr. 25, 2018. Available: https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe.

[7] T. Dutton, "An Overview of National AI Strategies," *Medium*, June 28, 2018. Available: https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd.

[8] R. Calo, The Case for a Federal Robotics Commission, Brooking Institute (2014). Available: https://www.brookings.edu/research/the-case-for-a-federal-robotics-commission/.

[9] G.E. Marchant. and K.W. Abbott, "International Harmonization of Nanotechnology Governance through "Soft Law" Approaches," *Nanotechnology Law & Business,* 9[4]: pp. 393-410, March 2013.

[10] S. A. Hetcher, "FTC as Internet Privacy Norm Entrepreneur," *Vanderbilt Law Review* 53[6]: pp. 2041-2062, 2000.

[11] G.E. Marchant, "'Soft Law' Mechanisms for Nanotechnology: Liability and Insurance Drivers," *Journal of Risk Research* 17: pp. 709-719, Feb. 14, 2014.

[12] J.W. Stempel, "Assessing the Coverage Carnage: Asbestos Liability and Insurance after Three Decades of Dispute," *Connecticut Insurance Law Journal* 12[2]: pp. 349-476 (2005).

[13] L. Fatehi et al. "Recommendations for Nanomedicine Human Subjects Research Oversight: An Evolutionary Approach for an Emerging Field," Journal of Law, Medicine & Ethics, 40(4): pp.716-750, Winter 2012.

[14] N. Bostrom, *Superintelligence.* New York, NY: Oxford University, Sept. 3, 2014.

[15] W. Wallach, and C. Allen, *Moral Machines: Teaching Robots Right From Wrong.* New York, NY: Oxford University Press, Nov. 19, 2008.

[16] M. Anderson, M. and S. L. Anderson, (eds.), Machine Ethics. New York, New York: Cambridge University Press, 2011.

[17] S. Russell, "Value Alignment," talk given at World Economic Forum meeting, Feb. 4, 2015. [Online]. Available: https://www.youtube.com/watch?v=WvmeTaFc_Qw

[18] Office for Human Research Protection, U.S. Department of Health and Human Services. International Compilation of Human Research Standards, 2017. [Online]. Available: https://www.hhs.gov/ohrp/sites/default/files/international-compilation-of-human-research-standards-2017.pdf

[19] C. Perrow, *Normal Accidents: Living With High-Risk Technologies*. New York: Basic, 1984).

[20] W. Wallach, “Building Global Infrastructure to Ensure AI and Robotics Are Beneficial,” Unpublished but widely circulated article, 2016.

[21] W. Wallach, “How to Keep AI from Slipping Beyond Our Control,” Geneva, Switzerland: The World Economic Forum (2017).

[22] K. Crawford and R. Calo, “There Is a Blind Spot in AI Research,” *Nature* 538: pp. 311-313 (2016).

"Soft Law" Governance of Artificial Intelligence

Gary Marchant

Introduction

On November 26, 2017, Elon Musk tweeted: "Got to regulate AI/robotics like we do food, drugs, aircraft & cars. Public risks require public oversight. Getting rid of the FAA wdn't (sic) make flying safer. They're there for good reason."[17]

In this and other recent pronouncements, Musk is calling for artificial intelligence (AI) to be regulated by traditional regulation, just as we regulate foods, drugs, aircraft and cars. Putting aside the quibble that food, drugs, aircraft and cars are each regulated very differently, these calls for regulation seem to envision one or more federal regulatory agencies adopted binding regulations to ensure the safety of AI. Musk is not alone in calling for "regulation" of AI, and some serious AI scholars and policymakers have likewise called for regulation of AI using traditional governmental regulatory approaches[18]

But these calls for regulation raise the questions of what aspects of AI should be regulated, how they should be regulated, and by who? The reality is that at best there will be some sporadic piecemeal traditional regulation of AI over the next few years, notwithstanding the increasing deployment and application of AI in a growing range of applications and industry sectors. In the interim at least, this "governance gap" for AI will mostly be filled by so-called "soft law" (see Part I, *supra*). These "soft law" mechanisms include various types of instruments that set forth substantive expectations but are not directly enforceable by government, and include approaches such as professional guidelines, private standards, codes of conduct, and best practices. A number of such soft law approaches have already been proposed or are being implemented for AI (see Part II, *supra*). While soft law has some serious deficiencies, such as lack of enforceability, there are additional strategies that can help maximize the effectiveness of this second-best approach to governance (see Part III, *supra*). For example, the lack of enforceability problem can be solved at least in part by various types of indirect enforcement by entities such as insurance companies, journal publishers, grant funders, and even governmental enforcement programs against unfair or deceptive business practices. Another problem, the lack of coordination between a potentially large number of overlapping and perhaps even inconsistent soft law programs, is to create what has been described as a Governance Coordinating Committee to help serve a coordinating function.

I. The Unsuitability of Traditional Regulation for AI

While some piecemeal regulation of specific AI applications and risks using traditional regulatory approaches may be feasible and even called for, AI has many of the characteristics of other emerging technologies that make them refractory to comprehensive regulatory solutions.[19] For example, AI involves applications that cross multiple industries, government agency

[17] https://twitter.com/elonmusk/status/934889932807593984?lang=en.

[18] *See, e.g.,* Paul Nemitz, *Constitutional Democracy and Technology in the Age of Artificial Intelligence*, 376 PHIL. TRANS R. SOC. A 20180089 (2018), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3234336; Frank Pasquale, *Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society*, 78 OHIO ST. L.J. 1243, 1252-55 (2017).

[19] Gary E. Marchant & Wendell Wallach, W. 2016. *Introduction.* In EMERGING TECHNOLOGIES: ETHICS, LAW AND GOVERNANCE, 1-12 (2016).

jurisdictions, and stakeholder groups, making a coordinated regulatory response difficulty. In addition, AI raises a wide range of issues and concerns that go beyond traditional regulatory agency focus on health, safety and environmental risks. Indeed, many risks created by AI are not within any existing regulatory agency's jurisdiction, including concerns such as technological unemployment, human-machine relationships, biased algorithms, and existential risks from future super-intelligence.

Moreover, the pace of development of AI far exceeds the capability of any traditional regulatory system to keep up, a challenge known as the "pacing problem" that affects many emerging technologies.[20] The risks, benefits and trajectories of AI are all highly uncertain, again making traditional preemptory regulatory decision-making difficult. And finally, national governments are reluctant to impede innovation in an emerging technology by preemptory regulation in an era of intense international competition.

For these reasons, it is safe to say there will be no comprehensive traditional regulation of AI for some time, except perhaps if some disaster occurs that triggers a drastic and no doubt poorly-matched regulatory response. Again, there may be slivers of the overall AI enterprise that are amenable to traditional regulatory responses, and these should certainly be pursued. But these isolated regulatory advances will be insufficient alone to deal with the safety, ethical, safety, military and existential risks posed by AI. Something more will be needed.

That something more that will be needed to fill the governance gap for AI will, at least in the short term, be within the category of "soft law." Soft law are instruments that set substantive expectations that are not directly enforceable by government. They can include private standards, voluntary programs, professional guidelines, codes of conduct, best practices, principles, public-private partnerships and certification programs. Soft law can even include what Wendell Wallach and I refer to as "process soft law" approaches such as coding machine ethics into AI systems or creating oversight systems within a corporate Board of Directors.[21] These types of measures are inherently imperfect, precisely because they are not directly enforceable.

This core weakness results in many other limitations, such that participation is incomplete, with the "good guys" complying and the "bad guys" not. These soft law measures are sometimes used as "whitewashing" (or "greenwashing") to make it look like a problem is being addressed when it really is not. And soft law measures are often expressed in vague, general language that is hard to measure compliance with. Finally, soft law measures generally do not provide the same reassurance to the public as traditional government regulation that the problems presented by a new technology are being adequately managed. This public reassurance effect is an important secondary function of regulation.

Notwithstanding these significant limitations, soft law has become a necessary and inevitable component of the governance framework for virtually all emerging technologies, including AI. Traditional regulatory systems cannot cope with the rapid pace, diverse applications, heterogeneous risks and concerns, and inherent uncertainties of emerging technologies. So although soft law measures are a second best solution, they are often the only

[20] Gary E. Marchant, *The Growing Gap between Emerging Technologies and the Law.* In GARY E. MARCHANT ET AL. THE GROWING GAP BETWEEN EMERGING TECHNOLOGIES AND LEGAL-ETHICAL OVERSIGHT: THE PACING PROBLEM 19-33 (2011).

[21] Wendell Wallach & Gary Marchant, *An Agile Ethical/Legal Model for the International and National Governance of AI and Robotics* (IEEE, in press, 2018).

game in town, at least initially. It recalls the quote attributed to Winston Churchill that "democracy is the worst form of government, except for all the others."[22]

Soft law has important advantages that explain its growing popularity and gap filling role. Soft law instruments can be adopted and revised relatively quickly, without having to go through the traditional bureaucratic rulemaking process of government. It is possible to experiment with several different soft law approaches simultaneously, indeed sometimes creating a problem of a proliferation of inconsistent private standards and other soft law instruments. They can sometimes create a cooperative rather than adversarial relationship among stakeholders. They are not bound by limited agency delegations of authority, and so can address any and all concerns raised by a technology. And because they are not adopted by a formal legal authority, they are not restricted to a specific legal jurisdiction, but can have international application.

II. Existing AI Soft Law Examples

We are already seeing the rapid infusion of soft law initiatives and proposals into the AI governance space. [23] Indeed, the likely first ever governance proposal for AI (at that time focused on robotics) was Isaac Asimov's three laws of robotics first published in 1942.[24] These "laws" were actually a form of soft law as they had no formal legal authority. More recently, an early entry into the AI soft law landscape was a "robot ethics charter" that the government of South Korea initiated in 2007, even though no final version of the ethics charter has ever been posted online.

Institute of Electrical and Electronic Engineers (IEEE)

Perhaps the most comprehensive soft law initiative for AI was launched in 2016 by the IEEE, one of the world's largest standard-setting and professional engineering society.[25] This initiative, entitled "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, is intended to "ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity."[26] The Initiative has two intended outputs. The first is a guide known as *Ethically Aligned Design*, which has now been published as draft versions I and II for public comments. Version II is a document that exceeds 250 pages and that addresses over 120 policy, legal and ethical issues associated with AI, with

[22] In fact, what Churchill actually said on the floor of the House of Commons was: "No one pretends that democracy is perfect or all-wise. Indeed it has been said that democracy is the worst form of Government except for all those other forms that have been tried from time to time.…" House of Commons, 11 November 1947, quoted in WINSTON CHURCHILL & RICHARD LANGWORTH, CHURCHILL BY HIMSELF: *THE DEFINITIVE COLLECTION OF QUOTATIONS* 574 (2008).

[23] Indeed, there has been such a proliferation of soft law programs and proposals for AI that the following examples provide just a sampling and not a comprehensive listing.

[24] Asimov's three laws are: (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) A robot must obey the orders given it by human beings except where such orders would conflict with the First Law; and (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws. These laws were first published in the 1942 short story *Roundabout,* which was published is Asimov's 1950 collection *I, Robot*.

[25] https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

[26] IEEE, Background, Mission and Activities of The IEEE Global Initiative, available at https://standards.ieee.org/develop/indconn/ec/ec_about_us.pdf.

recommendations assembled from more than 250 expert participants.[27] It seeks to "advance a public discussion about how we can establish ethical and social implementations for intelligent and autonomous systems and technologies, aligning them to defined values and ethical principles that prioritize human well-being in a given cultural context, inspire the creation of Standards (IEEE P7000™ series and beyond) and associated certification programs, [and] facilitate the emergence of national and global policies that align with these principles."[28] The final version of *Ethically Aligned Design* is scheduled to be published in 2019.

The second and even more relevant activity by the Initiative is to produce a series of IEEE standards addressing governance and ethical aspects of AI. The IEEE has given official approval to create the following standards, with standard-setting committees now established to develop each standard:

IEEE P7000™ - Model Process for Addressing Ethical Concerns During System Design
IEEE P7001™ - Transparency of Autonomous Systems
IEEE P7002™ - Data Privacy Process
IEEE P7003™ - Algorithmic Bias Considerations
IEEE P7004™ - Standard on Child and Student Data Governance
IEEE P7005™ - Standard for Transparent Employer Data Governance
IEEE P7006™ - Standard for Personal Data Artificial Intelligence (AI) Agent
IEEE P7007™ - Ontological Standard for Ethically Driven Robotics and Automation Systems
IEEE P7008™ - Standard for Ethically Driven Nudging for Robotic, Intelligent, and Automation Systems
IEEE P7009™ - Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
IEEE P7010™ - Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems

These ten AI standards are scheduled to be finalized by the end of 2021, and will provide a broad set of governance requirements relating to the governance of AI. For example, the chair of the working group developing standard IEEE P7006 on personal AI agents has recently written that the standard is being developed to provide "a principled and ethical basis for the development of a personal AI agent that will enable trusted access to personal data and increased human agency, as well as to articulate how data, access and permission can be granted to government, commercial or other actors and allow for technical flexibility, transparency and informed consensus for individuals."[29]

Partnership on AI

Another significant "soft law" player in the AI field is the Partnership on AI. This Partnership was originally started by the big players in the AI space such as Google, Microsoft, Facebook, IBM, Apple and Amazon, but has expanded to include a wide variety of companies,

[27] IEEE, Ethically Aligned Design Version II (2017), available at http://standards.ieee.org/develop/indconn/ec/auto_sys_form.html.

[28] *Id.* at ii.

[29] Katryna Dow and Marsali Hancock, *Injecting Ethical Considerations in Innovation Via Standards – Keeping Humans in the AI Loop,* IEEE Insight, Apr. 25, 2018, available at https://insight.ieeeusa.org/articles/standards-address-ai-ethical-considerations/.

think tanks, academic AI organizations, professional societies, and charitable groups such as the ACLU, Amnesty International, UNICEF and Human Rights Watch.[30] One of the stated goals of the Partnership is to develop and share best practices for AI which includes: "Support research, discussions, identification, sharing, and recommendation of best practices in the research, development, testing, and fielding of AI technologies. Address such areas as fairness and inclusivity, explanation and transparency, security and privacy, values and ethics, collaboration between people and AI systems, interoperability of systems, and of the trustworthiness, reliability, containment, safety, and robustness of the technology."[31]

The Partnership on AI has published a set of "Tenets" that include:

> "We are committed to open research and dialogue on the ethical, social, economic, and legal implications of AI….
>
> We believe that AI research and development efforts need to be actively engaged with and accountable to a broad range of stakeholders….
>
> We will work to maximize the benefits and address the potential challenges of AI technologies, by: Working to protect the privacy and security of individuals….Working to ensure that AI research and engineering communities remain socially responsible, sensitive, and engaged directly with the potential influences of AI technologies on wider society….Ensuring that AI research and technology is robust, reliable, trustworthy, and operates within secure constraints….Opposing development and use of AI technologies that would violate international conventions or human rights, and promoting safeguards and technologies that do no harm.
>
> We believe that it is important for the operation of AI systems to be understandable and interpretable by people, for purposes of explaining the technology.[32]

It remains to be seen if and how the Partnership will advance beyond these general tenets to produce more specific best practices and guidelines for responsible AI research and applications.

Future of Life Institute

The Future of Life Institute convened a meeting of many leading AI practitioners and experts at the Asilomar conference center in 2017, which is the home of the famous Asilomar Conference on Recombinant DNA held in 1975 which pioneered the soft law governance of technology by agreeing on a set of voluntary guidelines for genetic engineering research. At the 2017 Asilomar conference, the participants agreed on 23 principles to guide AI research and applications.[33] These principles include "Failure Transparency" ("If an AI system causes harm, it should be possible to ascertain why."); "Responsibility" ("Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.") and "Value Alignment" ("Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.").[34]

Industry groups have adopted their own soft law instruments for AI. For example, the Information Technology Industry Council (ITI) has developed its own set of AI principles.[35] For example, these principles include a commitment to "recognize our responsibility to integrate

[30] https://www.partners.org

[31] https://www.partnershiponai.org/#s-goals.

[32] https://www.partnershiponai.org/tenets/

[33] https://futureoflife.org/ai-principles/

[34] *Id.*

[35] https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf

principles into the design of AI technologies, beyond compliance with existing laws…. As an industry, it is our responsibility to recognize potentials for use and misuse, the implications of such actions, and the responsibility and opportunity to take steps to avoid the reasonably predictable misuse of this technology by committing to ethics by design."[36] The statement of principles, itself a form of soft law governance, also states a commitment to soft law principles: "We promote the development of global voluntary, industry-led, consensus-based standards and best practices. We encourage international collaboration in such activities to help accelerate adoption, promote competition, and enable the cost-effective introduction of AI technologies."[37]

Company-Specific Soft Law Initiatives

Some individual companies have also adopted their own statement of principles or guidelines for AI. For example, in June 2018 Google's CEO Sundar Pichai announced a set of seven principles that Google will follow in its AI activities.[38] Other major AI companies such as Microsoft[39] and IBM[40] have also announced their own AI principles that will guide their conduct.

Governmental AI Soft Law Initiatives

Governments have also supported the use of soft law methods to govern AI. The EU Commission published its strategy paper on AI on April 25, 2018.[41] Contrary to what many members of the European Parliament had hoped for and requested,[42] the Commission did not propose any new regulatory measures for AI at this time. Rather, it committed to develop a set of draft guidelines by the end of 2018.[43] However, the Commission did note that "[w]hile self-regulation can provide a first set of benchmarks against which emerging applications and outcomes can be assessed, public authorities must ensure that the regulatory frameworks for developing and using of AI technologies are in line with these values and fundamental rights. The Commission will monitor developments and, if necessary, review existing legal frameworks to better adapt them to specific challenges, in particular to ensure the respect of the Union's basic values and fundamental rights."[44]

Similarly, the UK House of Lords issued a detailed report on AI earlier in April 2018 and likewise recommended an ethical code of conduct for AI rather than any traditional "hard" regulation.[45] The report cited testimony on "the possible detrimental effect of premature

[36] *Id.* at 3.

[37] *Id.* at 5

[38] Sundar Pichai, AI at Google: Our Principles June 7, 2018), https://www.blog.google/technology/ai/ai-principles/.

[39] Microsoft, Microsoft AI Principles (undated), https://www.microsoft.com/en-us/ai/our-approach-to-ai.

[40] IBM, IBM's Principles for Trust and Transparency (May 30, 208), https://www.ibm.com/blogs/policy/trust-principles/.

[41] Communication From The Commission To The European Parliament, The European Council, The Council, The European Economic And Social Committee And The Committee Of The Regions: Artificial Intelligence for Europe {SWD(2018) 137 final}, April 25, 2018, available at file:///C:/Users/Gary/Downloads/CommunicationArtificialIntelligence.pdf.

[42] European Parliament resolution with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)); European Economic and Social Committee opinion on AI (INT/806-EESC-2016-05369-00-00-AC-TRA).

[43] Commission Communication, *supra* note 26, at 15.

[44] *Id.* at 16.

[45] UK House of Lords, Select Committee on Artificial Intelligence, Report of Session 2017–19, HL Paper 100

regulation" such as that "the pace of change in technology means that overly prescriptive or specific legislation struggles to keep pace and can almost be out of date by time it is enacted" and that lessons from regulating previous technologies suggested that a "strict and detailed legal requirements approach is unhelpful".[46] Based on such testimony, the House of Lords therefore concluded that "[b]lanket AI-specific regulation, at this stage, would be inappropriate."[47]

Instead, the House of Lords recommended a soft law strategy at least in the interim: "We recommend that a cross-sector ethical code of conduct, or 'AI code', suitable for implementation across public and private sector organisations which are developing or adopting AI, be drawn up and promoted … with a degree of urgency…. Such a code should include the need to have considered the establishment of ethical advisory boards in companies or organisations which are developing, or using, AI in their work. In time, the AI code could provide the basis for statutory regulation, if and when this is determined to be necessary."[48]

III. Evaluation and Moving Forward

A variety of entities from the government, industry and the non-government sectors have proposed or adopted soft law initiatives for the governance of AI. These soft law instruments include private standards, best practices, codes of conduct, principles and voluntary guidelines. They are in various states of development and implementation, and individually and collectively provide some initial guidance for the governance of AI. However, they suffer from major limitations. One prevalent problem is the generality of most of the provisions in these instruments. To some degree, this vagueness is inevitable and necessary, given the broad range of AI applications and the rapid pace and uncertain trajectory of its future, making precise requirements difficult if not impossible. Indeed, this is the very reason why the technology is primarily being governed by soft law rather than traditional hard law approaches at this time.

Two other limitations of the current matrix of soft law programs are however more amenable to progress and improvement. First, the unenforceability of these soft law provisions is the Achilles' heel of soft law approaches generally. There is no assurance or requirement that all, or even any, AI developers and users comply with the soft law recommendations. However, there are a number of mechanism that can be used to indirectly enforce these soft law provisions. Any entity with a supervisory role can adopt and monitor compliance with one or more AI soft law programs. For example, a corporation could create a committee of its Directors or a free-standing ethics committee and task it with ensuring compliance with the appropriate guidelines or codes of conduct adopted by or agreed to by that company. Universities could use the existing chain of authority, such as through department heads and deans, to require compliance with specified soft law AI provisions as part of the annual evaluation of faculty and staff. Or universities could create new, or expand the jurisdiction of existing, research oversight committees such as the Institutional Biosafety Committee to ensure adherence with specified AI soft law provisions.

Other actors could also play an important role in indirect enforcement of AI soft law programs. Certification bodies could create certification programs to certify that a company or other entity is adhering to a particular set of guidelines or principles. Business partners could

AI in the UK: Ready, Willing and Able? April 16, 2018, available at https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

[46] *Id.* at 113.

[47] *Id.* at 116.

[48] *Id.* at 125.

require certification with applicable AI soft law programs as a condition of doing business with that company. Insurers could require the implementation of appropriate AI risk management programs as a condition of liability coverage, just as some did with nanotechnology.[49] Granting agencies could condition funding on compliance with specified AI guidelines or codes of conduct. Professional journals could require compliance with certain best practices or guidelines as a condition of publication.

Even more legal quasi-enforcement approaches could be pursued. The Federal Trade Commission (FTC), under its general authority to take enforcement against deceptive and unfair business practices, could take enforcement action against a company that publicly commits to comply with a certain code of conduct or best practices but then fails to live up to its commitment. Private standards, especially those adopted by well-known standard setting bodies such as the IEEE, could be used to set a standard of care in tort law, and a company's failure to adhere to such standards, even though they are voluntary, could be evidence of failure to use reasonable care in a product liability or personal injury lawsuit.[50]

Soft law measures result in experience and field testing that can provide learning for subsequent traditional regulation. Indeed, soft law can sometimes be seen as a transitionary phase of governance that gradually "hardens" into traditional government regulation.[51] We may already be starting to see this hardening process of soft law in the AI space – for example, the State of California recently adopted legislation "expressing support" for the Asilomar AI Principles.[52]

Second, the confusing proliferation of different AI soft law programs and proposals creates confusion and overlap with regard to AI governance. It is hard for an actor in the AI space to assess and comply with all these different soft law requirements. Where do these various soft law programs overlap and duplicate each other? Where do they contradict each other? What gaps are not addressed by any of the existing soft law proposals? Some type of coordination is needed.

Wendell Wallach and I have proposed such a coordinating entity, which we have called a Governance Coordinating Committee (GCC).[53] This entity would not seek to duplicate or supplant the many organizations working on developing governance approaches to AI, but rather would provide a coordinating function much like an orchestra conductor in ensuring all the various players were connected with each other and aware of and responsive to each other's proposals, while also identifying gaps and inconsistencies in existing programs.

Conclusion

Soft law measures are very imperfect governance tools because of their lack of enforceability and accountability, as well as often being written in very general and self-serving language. Yet, for a rapidly developing and expansive technology like AI, comprehensive regulation by governments is not feasible, at least in the short term wit at best piecemeal

[49] Gary E. Marchant, *'Soft Law" Mechanisms for Nanotechnology: Liability and Insurance Drivers*. 17 J. RISK RESEARCH 709-719 (2014).

[50] *Id*.

[51] Gary E. Marchant, Douglas S. Sylvester & Kenneth W. Abbott, *Risk Management Principles for Nanotechnology,* 2 NANOETHICS 43, 53-54 (2008).

[52] State of California, Assembly Concurrent Resolution No. 215 (Sept. 7, 2018).

[53] Gary E. Marchant & Wendell Wallach, *Coordinating Technology Governance*, ISSUES IN SCIENCE & TECHNOLOGY, Summer 2015: 430-450.

regulatory enactments possible. Accordingly, soft law will be the default approach for most AI governance at the present time. For that reason, there is a need to explore ways to indirectly enforce and coordinate the proliferation of soft law measures that have already been proposed or enacted for AI.

Section VI. Conclusion

The Hastings Center AI workshops played a seminal role in catalyzing the creation of a transdisciplinary community of scholars dedicated to the safety and ethics of artificial intelligence and robotics. By way of illustration, this report will highlight how the profile and contributions of just a few project participants were greatly expanded from the first workshop to the present.

Francesca Rossi, a participant in all three workshops, the IBM AI Ethics Global Leader and distinguished research scientist at the IBM T.J. Watson Research Center, is a central figure in the emergence of an array of projects within the AI safety and ethics community. She is a founding board member for the Partnership on AI, a past president of the International Joint Conference on Artificial Intelligence (IJCAI), sits on the European Commission's High-Level Expert Group on Artificial Intelligence, was a member of The World Economic Forum's Global Future Council on Artificial Intelligence and Robotics, and sits on the executive committee of the Institute of Electrical and Electronics Engineers (IEEE) global AI initiative. Of particular note is Francesca's role in proposing and organizing a new AI, Ethics, and Society (AIES) conference co-sponsored by two of the leading professional bodies, the Association for the Advancement of Artificial Intelligence (AAAI) and the Association for Computing Machinery (ACM). The motivation for this conference was the need for a multidisciplinary scientific conference for peer-reviewed papers. AI research, AI and law, AI and economics, and AI and philosophy were among the topics covered in the 62 papers, two panels, and four invited talks at the first AIES, co-located with the AAAI annual conference in New Orleans in 2018. The proceedings were published in the ACM and the AAAI Digital Library after the conference. A second conference will be held in 2019, and will again be co-located with the AAAI annual conference.

AIES 2018: http://www.aies-conference.com/2018/
AIES 2019: http://www.aies-conference.com/

John Havens, a participant in workshops I and III, had, at the time of the first Hastings AI workshop, just published his book *Heartificial Intelligence* and had been appointed Executive Director of the IEEE Global Initiative on Ethics of Autonomous Systems. The IEEE (Institute of Electrical and Electronics Engineers) is a professional association noted for establishing international standards. John writes that, at the first Hastings workshop, he "learned a great deal about the technical and legal landscape of AI ethics" and met Illah Nourbakhsh, Anja Kaspersen, Yann LeCun, and others who contributed to his understanding and to the IEEE Initiative. Under the leadership of John and Konstantinos Karachalios (Managing Director of IEEE Standards Association), the IEEE Global Initiative has grown to be one of the most important projects for establishing AI and robotic norms, standards, and best practices. The project now includes more than 1000 individuals, and more than 200 of these contributed to either the first version of *Ethically Aligned Design* (EAD 2016) or *Ethically Aligned Design version 2* (December 2017). EAD inspired the formation of 14 Standards Working Groups under the IEEE P7000 Series, as well as other important activities. Many members of the Hastings workshops participate in these standards-setting committees. The IEEE Global Initiative on Ethics of Autonomous Systems will continue to grow as one of the most significant international initiatives promoting standards and guidelines for safe and ethical AI.

Anja Kaspersen, a participant in all three workshops, attended the first workshop representing the World Economic Forum, where she spearheaded the Forum's work on geopolitics, international security, and new technologies. She attended the second workshop as the Head, Strategic Engagement and New Technologies for the International Committee of the Red Cross. By the time of the third workshop, she was Director of Disarmament Affairs at the UN in Geneva. She also co-founded Building Global Infrastructure for the Governance of AI (BGI4AI) with Wendell Wallach. Anja is highly regarded as a diplomat and a leading expert on the impact of emerging technologies, particularly AI, on international security. She credits the workshop and the network of relationships she forged at these gatherings as playing a significant role in her understanding and the evolution of her thought. She has also been a champion for the advancement of many skilled women leaders, and even helped some she met at Hastings workshops in furthering their careers.

Kay Firth-Butterfield (workshop I) and **Erika Kochi** (workshop III): At the time of the first workshop Kay Firth-Butterfield, a Barrister, was the Chief Office of the Ethics Advisory Panel for LUCID Technologies, Inc. She is now Head, Artificial Intelligence and Machine Learning for the World Economic Forum (WEF) Center for the 4th Industrial Revolution in San Francisco. During the third Hastings workshop an extended discussion formulated specific guidelines for children and AI. The discussion contributed toward inspiring Erica Kochi, the co-founder of UNICEF Innovation, to create with Kay Firth-Butterfield a UNICEF/WEF Children and AI initiative.

With the increasing importance and attention given to AI many of the workshop participants have been busy speaking about research or the safety and ethics of AI at conferences and other gatherings. Particularly prominent have been the continual travels of Francesca Rossi, Stuart Russell, and Wendell Wallach to events around the globe. They have become part of a small cadre loosely weaving the various institutions entering this space together, helping to enhance cooperation, and to defuse competition.

Recommendations

Two recommendations emerging from the workshops have already been discussed in this report.

1) A consortium of industry leaders, international governmental bodies, and nongovernmental institutions, national and regional (such as 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 turned to 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 NYC 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.

A third recommendation states:

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. AI and health care is 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.

Let us unpack this third recommendation a bit.

Several industry-led initiatives are under way to explore the ethical and social implications of AI, including work by IEEE and PAI, among others. In general, these initiatives have adopted a horizontal approach to the issues, with work groups organized around specific topic areas, such as safety, transparency, fairness, labor force implications, and the knock-on impact of AI on human-AI relationships. This horizontal approach has been valuable in surfacing important values and tradeoffs at stake. Use-cases drawing on experience across society and the economy have provided meaningful insights to guide creation of industry standards and practices.

To complement these horizontal initiatives, a project should be commissioned to conduct a deep-dive into all issues pertinent to the U.S. healthcare system, comprised of health care research, clinical practice, and public health. The project would recognize lessons and insights being developed in existing horizontal initiatives but would leverage them as a platform for further exploration at a deeper level. Importantly, this approach also would seek to explore AI-related implications specific to health care that are "out of scope" for issue-specific working groups.

This vertical approach offers several advantages:

- Urgency. The health care sector, broadly defined, is a top-priority investment opportunity for the application of AI, with target areas including drug R&D, genomics, precision medicine, clinical trials, patient record management, disease diagnostics and treatment, assisted surgery, and epidemic management, to name a few. Competition to capture new market share is intense, and deployment often precedes full consideration of consequences.

- Focus. This would ensure that all issues, implications, and tradeoffs pertinent to a particular sector of U.S. society and economy come to the fore, including those that may be overlooked, underemphasized, or drop through the cracks in a horizontal attack of the issues.

- Perspective. This approach would facilitate meaningful inclusion of constituents and stakeholders who might otherwise be ignored in a horizontal approach to the issues, including doctors, nurses, patients (via advocacy groups), hospital administrators, pharmaceutical executives, and public health officials.

This is a significant undertaking, and it might best be accomplished as a series of related projects, either as sub-projects, or conducted in phases over time. For example, an initial project that begins to map the ethical and trade-off concerns of healthcare in a broad but cursory manner, could be followed by workshops that deep-dive individually into research, clinical practice, public health, and other areas of interest.

*   *   *

The Hastings AI and robotics workshops may or may not be remembered when a history about the evolution of safety and ethics for autonomous systems is written. Nevertheless, these engagements of key figures within an emerging field can play a significant role in shaping that field and in shaping the understanding and future activities of those leaders in an unfolding story. While much of the influence of well-designed transdisciplinary conversations is hard to capture and may not even be fully recognized by participants whose insights and understanding have been sparked, The Hastings Center is pleased that these workshops also produced the tangible results outlined in this report.

Appendix: Project Participants

Colin Allen
Distinguished Professor, Department of History and Philosophy of Science
Faculty, Center for the Neural Basis of Cognition
University of Pittsburgh

Ronald C. Arkin
Regents' Professor
Director, Mobile Robot Laboratory
School of Interactive Computing
Georgia Tech

Amir Banifatemi
General Manager, Innovation & Growth
Executive Director, AI XPRIZE and ANA Avatar XPRIZE
XPRIZE Foundation

Brandon M. Belford
Government Affairs
Apple, Inc.
Former Senior Policy Advisor, National Economic Council

Selmer Bringsjord
Professor of Cognitive Science
Professor of Computer Science
Professor of Logic and Philosophy
Director, Rensselaer AI and Reasoning Laboratory
Rensselaer Polytechnic Institute (RPI)

Kay Firth-Butterfield
Head, Artificial Intelligence and Machine Learning
World Economic Forum LLC

David Chalmers
Professor of Philosophy and Neural Science
Co-Director, Center for Mind, Brain and Consciousness
New York University

Vincent Conitzer
Kimberly J. Jenkins University Professor
of New Technologies
Professor of Computer Science
Duke University
Andrew Critch
Research Scientist, Center for Human-Compatible Artificial Intelligence
Executive Director, Berkeley Existential Risk Initiative
University of California Berkeley

Mary "Missy" Cummings
Professor, Department of Mechanical Engineering and Materials Science
Professor of Computer Science
Director, Humans and Autonomy Lab
Duke University

Daniel Dewey
Program Officer, Potential Risks from Advanced Artificial Intelligence
Open Philanthropy Project

Thomas G. Dietterich
Distinguished Professor (Emeritus) of Computer Science
Institute for Collaborative Robotics and Intelligent Systems
Oregon State University
(President Emeritus of AAAI)

Anca Dragan
Assistant Professor, Electrical Engineering and Computer Science Department
Founder and Director, InterACT Lab
University of California Berkley

Pascale Fung
Professor, Department of Electronic and Computer Engineering
Director, Center for Artificial Intelligence Research (CAiRE)
Hong Kong University of Science and Technology

Amandeep Singh Gill, PhD
Co-Executive Director, Secretariat of the High-Level Panel on Digital Cooperation
Ambassador and Permanent Representative of India, Conference on Disarmament
United Nations

Tom Gruber
Head of Advanced Development Group
Apple, Inc.

Gillian K. Hadfield
Professor of Law
Professor of Strategic Management
University of Toronto
Senior Policy Advisor, OpenAI

Dylan Hadfield-Menell
Ph.D. student
University of California, Berkley

Josh Storrs Hall
Founding Chief Scientist, Nanorex, Inc.
Research Fellow
Institute for Molecular Manufacturing

Joe Halpern
Professor of Computer Science
Cornell University

John C. Havens
Executive Director, IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
Executive Director, Council on Extended Intelligence (CXI)
IEEE Standards Association
Deborah G. Johnson
Anne Shirley Carter Olsson Professor of Applied Ethics (Emeritus)
University of Virginia

Gregory Kaebnick
Editor-in-Chief, *The Hastings Center Report*
Research Scholar
The Hastings Center

Daniel Kahneman
Eugene Higgins Professor of Psychology
Professor of Psychology and Public Affairs (Emeritus)
Woodrow Wilson School, Princeton U.

Subbarao Kambhampati
Professor of Computer Science and Engineering
Arizona State University
Former President, AAAI

Behzad Kamgar-Parsi
Program Officer
Office of Naval Research

Anja Kaspersen
Director, Office for Disarmament Affairs
United Nations, Geneva

Erica Kochi
Co-founder
UNICEF Innovation

Yann LeCun
VP and Chief AI Scientist
Facebook
Founding Director, Center for Data Science
Silver Professor of Computer Science
Professor of Neural Science
Professor of Electrical and Computer Engineering
New York University

Sean Legassick
Co-Lead, Ethics & Society
DeepMind

Michael Littman
Co-Director, Humanity Centered Robotics Initiative (HCRI)
Professor of Computer Science
Brown University

Alan Mackworth
Professor of Computer Science
Founding Director, Laboratory for Computational Intelligence
University of British Columbia
Former President, AAAI
Former President, AJCAI

Gary Marchant
Regents' Professor
Sandra Day O'Connor College of Law
Founding Director, Center for Law and Innovation
Arizona State University

Gary Marcus
Founder, Geometric Intelligence
Professor of Psychology
New York University

Illah R. Nourbakhsh
K&L Gates Professor of Ethics and
Computational Technologies
The Robotics Institute
Carnegie Mellon University

Laurent Orseau
Research Scientist
Google DeepMind

David Roscoe
Chair, Advisory Council
The Hastings Center

Francesca Rossi
Global leader, IBM AI Ethics
Distinguished Research Scientist
IBM T.J. Watson Research Center
Professor of Computer Science
University of Padova, Italy
Former President, IJCAI

Stuart Russell
Professor of Computer Science
Michael H. Smith and Lotfi A. Zadeh Chair in
Engineering, Computer Science Division
University of California at Berkeley

Bernhard Schölkopf
Director, Empirical Inference
Max Planck Institute for Intelligent Systems
Bart Selman
Professor of Computer Science
Cornell University
President-Elect, AAAI

Nate Soares
Executive Director
Machine Intelligence Research Institute

Mildred Z. Solomon
President & CEO
The Hastings Center

John Sullins
Professor, Philosophy
Sonoma State University

Shannon Vallor
William J. Rewak, S.J. Professor of Philosophy
Santa Clara University

Wendell Wallach
Consultant, Ethicist, and Scholar
Interdisciplinary Center for Bioethics
Yale University
Senior Advisor, The Hastings Center

Toby Walsh
Professor of Artificial Intelligence
University of New South Wales, Australia

Daniel Weld
Thomas J. Cable/WRF Professor of Computer
Science and Engineering
University of Washington

Michael Wellman
Lynn A. Conway Professor of Computer Science
and Engineering
University of Michigan

David Woods
Professor of Cognitive Systems Engineering and
Human Systems Integration
Ohio State University