Responsible Governance of Artificial Intelligence:

An Assessment, Theoretical Framework, and Exploration

by

Miles Brundage

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2018 by the
Graduate Supervisory Committee:

David Guston, Chair
Erik Fisher
Lauren Keeler
Joanna Bryson

ARIZONA STATE UNIVERSITY

December 2019

ABSTRACT

While artificial intelligence (AI) has seen enormous technical progress in recent years, less progress has occurred in understanding the governance issues raised by AI. In this dissertation, I make four contributions to the study and practice of AI governance. First, I connect AI to the literature and practices of responsible research and innovation (RRI) and explore their applicability to AI governance. I focus in particular on AI's status as a general purpose technology (GPT), and suggest some of the distinctive challenges for RRI in this context such as the critical importance of publication norms in AI and the need for coordination. Second, I provide an assessment of existing AI governance efforts from an RRI perspective, synthesizing for the first time a wide range of literatures on AI governance and highlighting several limitations of extant efforts. This assessment helps identify areas for methodological exploration. Third, I explore, through several short case studies, the value of three different RRI-inspired methods for making AI governance more anticipatory and reflexive: expert elicitation, scenario planning, and formal modeling. In each case, I explain why these particular methods were deployed, what they produced, and what lessons can be learned for improving the governance of AI in the future. I find that RRI-inspired methods have substantial potential in the context of AI, and early utility to the GPT-oriented perspective on what RRI in AI entails. Finally, I describe several areas for future work that would put RRI in AI on a sounder footing.

ACKNOWLEDGMENTS

I am deeply indebted to numerous people for providing feedback on my work over the past few years, including especially my committee: David Guston, Erik Fisher, Lauren Keeler, and Joanna Bryson. I am grateful for their invaluable insights, and their flexibility with my completion of this dissertation abroad. I also appreciate the flexibility of my colleagues at the Future of Humanity Institute, especially Andrew Snyder-Beattie, Allan Dafoe, Nick Bostrom, and Carrick Flynn, who provided a stimulating work environment and were very understanding of my need to juggle different research projects. There are countless others who contributed valuable insights to this work, and whom I will be unable to list comprehensively, but I will briefly mention in particular, Shahar Avin, my primary collaborator on the workshop and report on malicious uses of artificial intelligence; David Kristoffersson, who assisted greatly on the engineering side of one of the formal modeling projects described herein; Baobao Zhang, who provided useful feedback on survey design; and in Arizona, Clark Miller, who helped at various points to move my PhD along and was always supportive of my work on AI, and Andra Williams, who tirelessly supported me and other students throughout my time at ASU. Finally, I am grateful to my family and friends for encouragement throughout this process, and my partner, Larissa Schiavo, for helping with final edits. My dissertation is dedicated to my parents, Dianne and Jan Brundage, who, in addition to being wonderful parents, have been particularly enthusiastic supporters of my research ambitions and my transition to working on AI.

TABLE OF CONTENTS

Page

LIST OF FIGURES …………………………………………………………………vi

CHAPTER

1: INTRODUCTION…………………………………………………………..1

2: A FRAMEWORK FOR RESPONSIBLE INNOVATION IN AI ………………….8

Preliminaries 8

Overview of Responsible Research and Innovation 11

AI’s Governance-Related Characteristics 17

The Importance of Publication Norms in RRI for AI 28

The Importance of Cooperation in RRI for AI 31

Conclusion 33

3: AN ASSESSMENT OF EXISTING EFFORTS ……………………………..34

Introduction.. 34

Overview of AI Governance 34

Anticipation.. 42

Inclusion.. 44

Reflexivity.. 47

Responsiveness 48

CHAPTER Page

Conclusion.. 51

4: A PRESCRIPTION: DISCIPLINED, REFLEXIVE AI FUTURES..................52

A Proliferation of AI Futures.. 52

Anticipation and Reflexivity... 53

The Comparison to Energy 55

Anticipatory Failures 57

Expert Elicitation, Scenario Planning, and Formal Modeling.............................. 59

Conclusion 61

5: SCENARIO PLANNING..63

Introduction.. 63

The Contestedness of AI Futures... 63

Scenario Planning: A Tool for Anticipation and Reflexivity Under Uncertainty 66

June 2017 Scenario Planning Workshop 68

Lessons Learned.. 74

Conclusion 79

6: EXPERT ELICITATION..80

Introduction.. 80

Expert Elicitation 80

CHAPTER Page

Scenario Planning Workshop 88

Lessons Learned 91

Conclusion 93

7: FORMAL MODELING …………………………………………………………...94

Introduction 94

Formal Modeling 94

Modeling AI Openness in an Agent-Based Modeling Framework 95

Game Theory as a Lens on an “AI Arms Race” 100

Conclusion 107

8: CONCLUSION AND FUTURE DIRECTIONS…………………………………108

Introduction 108

Contributions of the Analytical Framework 108

Toward an Integrated Portfolio of Anticipatory Methods 110

Normative Recommendations 113

Future Directions 115

BIBLIOGRAPHY………………………………………………………………………118

APPENDIX

A: SURVEY QUESTIONS…………………………………………………………..137

B: INSTITUTIONAL REVIEW BOARD DOCUMENTATION…...…..……..141

LIST OF FIGURES

Figure Page

1. Scenario Dimension Brainstorming Phase of Workshop……………….....................71

2. Notes on the 2x2 Scenario Matrix from Workshop………………………………….73

3. Excerpts from Scenario Sketches from Scenario Planning Workshop...…………….74

4. Table 1……………………………………..……………………..…………….91

5. Table 2……………………..………………………………..……………………..92

6. Generic 2x2 Game Matrix………………………………………………………104

CHAPTER 1: INTRODUCTION

Artificial intelligence (AI) has long been a subject of interest to researchers, entrepreneurs, science fiction writers, and the general public. There have been ebbs and flows of attention paid to the field since the mid-twentieth century. AI is currently receiving an unprecedented scale of attention, as shown, e.g., by analysis of New York Times coverage (Fast and Horvitz, 2016). Since the field's formal beginnings in the mid-20th century, and especially in recent decades, AI has sparked discussion of ethics and governance, and such discussions have also increased in recent years, as discussed further below.

Since roughly 2012, two parallel and related trends have transformed contemporary discussions of AI. First, deep learning, the training of neural networks with multiple hidden layers, has enabled a new wave of human-competitive performance on a wide array of diverse tasks, including image recognition, machine translation, and speech recognition (LeCun et al., 2015). This development builds on a longer history of Internet-related technologies (e.g., search and advertisement placing) leveraging AI techniques, but the application of AI to commercial purposes has received greater attention and enthusiasm in recent years. Second, a substantial increase in discussion of the societal implications of AI has taken place, with participants at different times calling for greater research, formal regulation, and/or some form of self-regulation from the AI community (Brundage and Guston, 2019). Partly inspired by the machine learning "revolution" and the associated growth in commercially and societally impactful applications of AI, these governance discussions have taken place in a wide range of countries and have reached the highest levels of government, including public remarks by former President Obama during his final year in office, President Xi Jinping, President Vladimir Putin, Prime Minister Justin Trudeau, President Emmanuel Macron, European Council President Donald Tusk, and Prime Minister Boris Johnson, among others.

The rising attention to governance within and around the field of AI has been characterized as a “scientific-intellectual movement” (Frickel and Gross, 2005) by Brundage and Guston (2019) but as yet has not been the subject of much thoughtful self-reflection by practitioners with regard to the ideal methods and frameworks applicable to AI governance. Indeed, a range of perspectives have been espoused, from those focusing governance concerns on contemporary AI systems and their associated social and economic implications (Elish et al., 2016; Campolo et al., 2017) to those focusing on hypothesized long-term safety risks associated with more powerful AI systems (Bostrom, 2014; Sutskever and Amodei, 2017) and with various perspectives that straddle or attempt to reconcile such apparently competing time horizons and foci (Executive Office of the President, 2016; Baum, 2017; Krakovna, 2018).

Despite this increasing attention, there is little agreement about the nature of the challenge to be solved in AI governance, and little connection of nascent AI governance efforts to existing scholarly frameworks for reasoning about scientific and technological governance.

Given this complex and rapidly evolving landscape of governance concerns, my dissertation aims to contribute to the more disciplined study and practice of AI governance. Specifically, I motivate, describe, and apply the responsible research and innovation (RRI) framework to the context of AI governance. I suggest that successfully delivering on RRI in this context demands attention to AI's status as a general purpose technology (GPT), and the implications that that status has for responsible publication and cooperation among AI developers. I draw on recent developments in the field to illustrate the practical relevance of my proposed framework.[1] Next, I motivate a set of methods for analyzing AI futures more rigorously and reflexively that match the nature of the AI governance problem well, and report on the results of initial efforts in this direction. Finally, I suggest areas for future work. Overall, I found that RRI is a productive framework for AI and that the specific methodological directions I pursue have potential to improve AI governance. In the remainder of this introduction, I briefly preview each of these contributions, which correspond to the order of the following chapters.

In **Chapter 2**, I give definitions of key terms and analyze AI's governance-related properties, including its status as a GPT. I briefly describe the history and evolution of thinking on the governance of science and technology, with a focus on the contemporary framework of RRI as my key touchstone in this evolution. I argue that RRI provides a rich framework for thinking about responsibility in AI, but that the challenges of publishing general purpose AI systems[2] have been neglected. Yet these issues loom large in the contemporary challenge of responsible research and innovation in AI. Further, the

[1] In several cases, I have been directly involved in the events I describe, especially with regard to the publication of the GPT-2 system.

[2] I discuss the meaning of generality in more detail later, but briefly I consider there to be a spectrum of increasing generality in which a technology can have more of an ability to be steered toward performing a diverse set of tasks with less human intervention required for each marginal additional task, compared to less general technologies (including in some cases earlier versions of the same AI system). For example, the language model GPT-2 is able to more efficiently adapt to new domains than earlier language models, and the larger versions of the system encapsulate more transferrable knowledge than the smaller versions. Radford and Wu et al, 2019; Solaiman et al., 2019; Brundage et al., 2019.

GPT framing puts a high premium on anticipation of the progress in underlying AI capabilities and the malicious uses toward which AI can be put, which motivates some of the anticipatory efforts described later.

In **Chapter 3**, I analyze recent developments in AI governance from the perspective of RRI. First, I note the long roots of reflection on societal impacts of AI, and then describe recent developments, especially those occurring in the past few years. I critique these developments with reference to the four “dimensions” of RRI developed by Stilgoe et al. (2013), namely anticipation, reflexivity, inclusion, and responsiveness. I find that anticipatory efforts in AI have generally been underdeveloped, reflexivity in the AI community is too low, inclusion has been narrowly construed, and that responsiveness to surfaced normative considerations has been minimal. This critical assessment provides a baseline for the methodological interventions I discuss in later chapters, which are targeted at filling gaps in the current landscape.

In **Chapter 4**, I delve into the question of methodology in AI governance analysis: given the RRI framework and the aforementioned characteristics of AI, how ought one go about shaping AI’s development and broader social context positively? I explain why my methodological exploration places a particular focus on the RRI dimensions of anticipation and reflexivity. AI futures are currently highly contested and ill-explored, and both experts and non-experts are insufficiently reflective about the risks, opportunities, and options potentially facing them. These challenges are exacerbated by the generality of AI systems. I contrast AI with energy in various respects. Energy is a technological domain in which anticipatory methods are better developed and more pervasively used, although one in which anticipatory challenges remain. I focus attention on identifying unstated assumptions about possible AI futures, the development of a common language for discussing alternative futures, and enabling more robust decision-making in the face of irreducible uncertainties (Lempert et al., 2003). Toward this end, I briefly explain and justify three methods for more disciplined future thinking in AI--expert elicitation, scenario planning, and formal modeling--which can aid in this reflexive anticipation. The following chapters explore each in more detail. I also note the importance of additional methodological reflection and action related to the dimensions of responsiveness and inclusion, which are less extensively explored in this dissertation.

In **Chapter 5,** I motivate and explore the use of scenario planning for AI governance analysis, and note the limits of related efforts to date. I then discuss a scenario planning exercise conducted in June 2017, co-facilitated with Lauren Keeler. I highlight the tight match between the characteristics of problems for which scenario planning was designed and the characteristics of AI governance. I report key lessons learned from this process, including the surfacing of previously conflated dimensions of possible futures and the highlighting of issues not well addressed in contemporary AI policy analyses. These are suggestive of the value of future scenario exercises related to AI governance. I highlight areas for possible improvement in future scenario planning exercises and feedback from participants on the value of the event for stimulating reflection on different futures.

In **Chapter 6**, I describe two cases of expert elicitation (conducted in February 2017 and June 2017), both of which show the potential for improving anticipatory capabilities and reflexive awareness among AI experts. In each case, I describe the methods employed and the resulting analysis, with an eye toward informing more frequent and pervasive uses of expert elicitation in the future. I also discuss the limitations of expert elicitation more generally, given the uncertainties related to technical trends in AI as well as uncertainty about what sorts of expertise are most relevant to governing a GPT. Additionally, I present novel findings related to normative disagreement among AI experts. Finally, I discuss recent collaborative work involving explicit anticipation of possible beneficial and harmful uses of a particular AI system (GPT-2) in order to inform responsible publication decisions.

In **Chapter 7**, I motivate the use of formal models of AI futures and describe two cases of applying such methods to AI. First, I describe an early effort to design an agent-based model of AI futures, the practical and conceptual challenges of which are informative. The model attempted to capture some salient properties of openness in AI. I describe the design decisions that went into the model's characterization of openness, which involved issues not (to my knowledge) previously analyzed in the literature, such as the distribution of resources and the absorption rate of shared AI results. In the second case, I describe an effort to reframe the emerging narrative of an "AI arms race" in explicit game theoretic terms, posing the question of whether such a race, if one exists, is best thought of as a Prisoner's Dilemma, a Stag Hunt, or some other canonical "game" (in the sense used in game theory). I report positive feedback related to this framing and highlight aspects of AI governance for further study which were surfaced as a result of taking this modeling approach. I also discuss more recent collaborative research in which I was involved that pushes this line of thinking further toward identifying concrete policy implications. This work outlines a coherent framework for thinking about the solution of collective action problems in AI--problems made more severe by the generality of AI systems. The apparent returns on a modest investment in modeling AI development suggests that there is insufficient reflexivity today regarding “hot” AI governance topics such as openness, and that formalization can be one means of increasing reflexivity.

In **Chapter 8**, I distill lessons learned and future directions from the above analysis. I describe synergies between the methods described in earlier chapters, and identify ways to perform these methods better in the future. I make two recommendations for those involved in governing AI: first, a more systematic effort to identify opportunities in the broad area of “AI for good,” a particularly promising possibility afforded by AI’s status as a GPT; and second, increased attention to inclusion in discussions of AI’s future. Absent such democratization of foresight for and shaping of AI, the anticipatory tools described here might be used to entrench power rather than to steer AI in broadly beneficial directions. Finally, I discuss several areas for future work, including scaling up the methods described, deepening them in various respects, and improving the theoretical foundation of AI governance through comparative analyses of other general purpose technologies.

CHAPTER 2: A FRAMEWORK FOR RESPONSIBLE INNOVATION IN AI

Introduction

In this chapter, I develop a general account of AI as a subject of governance and situate my proposed analytical framework in the context of other frameworks, especially responsible research and innovation (RRI).[3] First, I provide definitions of AI, responsible, and governance for the purpose of my analysis. Next, I elucidate the recent history of thinking on the governance of emerging technologies, focusing in particular on the growing area of RRI. Then I describe the governance-related characteristics and social context of AI, with the aim of identifying distinctive challenges of RRI in that context. I focus in particular on AI's status as a general purpose technology (GPT), the unequal distribution of inputs to its development, and its scalability. These features of AI motivate the need for a greater attention to decisions surrounding the publication of general purpose systems and for cooperation between actors in order for AI development to be responsible. This GPT-oriented perspective on RRI's practical implications for AI, along with my assessment of extant efforts in the next chapter, motivate my emphasis on methods for anticipation and reflection in the remainder of the dissertation.

Preliminaries

By artificial intelligence (AI), I mean digital systems that respond appropriately to uncertain opportunities and affordances in their environment, often in part through

[3] Note that responsible innovation is often used as shorthand for responsible research and innovation, as in the title of this chapter.

learning.[4] The "environment" in question might be fully digital, such as a dataset that needs to be labeled, or external, such as the immediately surrounding physical world in the case of an AI-enabled robot. Representative examples of AI systems include search engines, speech recognition systems, semi-autonomous drones, and machine translation systems. The term "AI" has been used to refer to various things, including a research community with the long-term ambition of building more broadly competent digital systems, or the specific technical artifacts already produced by that field, or the systems which researchers in the aforementioned field might aspire to build in the future. Each of these definitions is relevant to questions of governance in different ways: the field of AI is having unprecedented economic and social influence today, many AI systems are already having an impact on society (Brundage and Bryson, 2014; Brundage and Bryson, 2016), and future technical and social developments should inform the nature and degree of our concern about and societal preparation for them (Brundage, 2015; Brundage, 2016a). While each of these is relevant to governance, I primarily focus on issues related to the design and dissemination of AI systems, in which the AI community is a key actor. When I refer to AI being governed, I am referring to the set of institutions, norms, and laws surrounding digital systems that are intended to have some degree of "intelligence," regardless of whether that intelligence is explicitly modeled on biological organisms such as humans or not.

The use of the term "uncertain" in my definition distinguishes AI from other, more static information and communication technologies (ICTs) which are technologies

[4] This definition is adapted from Brundage and Bryson, 2016.

that have full information about their tasks, act deterministically on fully structured data, and cannot be well described as acting in the pursuit of goals. AI systems are agents in the sense that they pursue goals (Russell and Norvig, 2009), but this does not imply human-likeness. AI systems are designed as artifacts to flexibly achieve goals using limited information and computational capacity.

"AI system," as I use the term in this dissertation, can be thought of as shorthand: "AI socio-technical system" would in many cases be a better reflection of the designed[5] and socially embedded nature of the artifacts built by the AI community.[6] Also note that, while much is often made of the definition of AI in popular culture and some policy discussions, the subsequent discussions do not hinge greatly on my definition being used versus another. An alternative definitional cluster focuses on digital systems that perform some behavior which, if done by humans or non-human animals, would be seen as requiring intelligence. Both the generic definition I use and one based on reference to human and non-human animal behavior would yield similar (though not identical) conclusions about the pervasive social and economic applications and implications of AI. A technology capable of substituting for either a subset of "intelligence" generally or a subset of "human-like intelligence" specifically would both have an enormous range of applications, even though these are technically distinct. The definition I use avoids

[5] Many AI systems learn from their experience, but this does not negate the fact that people make a range of design choices when creating and operating them.

[6] Importantly, referring to "AI systems" as agents of societal change does not imply moral agency on the part of the technology. Bryson (2019) notes that "no fact of either biology (the study of life) nor computer science (the study of what is computable) names a necessary point at which human responsibility should end. Responsibility is not a fact of nature. Rather, the problem of governance is as always to design our artefacts—including the law itself—in a way that helps us maintain enough social order so that we can sustain human flourishing."

unnecessary anthropomorphism: AI systems needn't be (and in practice only sometimes are) modeled after biological systems, and a definition anchored in humans as the gold standard for intelligence can be misleading.[7] Finally, note that I use the term AI to refer both to existing systems as well as those that are plausible in the future, contra those who would describe existing systems as not constituting "real" AI (Brundage and Bryson, 2014).

By responsible, I mean that an actor takes ownership for the consequences of their actions, is mindful of possible alternative actions to take, and acts in a way that is cognizant of their ethical, legal, social, and other obligations.

By governance, I mean authoritative human decision-making related to a topic, domain, artifact, or political jurisdiction. Governance can be "hard" (e.g., enforceable laws), "soft" (e.g., codes of conduct for which only social opprobrium results for non-compliance), centralized, decentralized, public, and private. Strong governance of AI is needed to ensure continued human accountability for the impacts of AI systems (Bryson, 2018), and there is a flowering literature on the need for robust AI governance (Calo, 2017). I turn now to a key foundation of my approach to responsibly governing AI: the intellectual and practical tradition of responsible research and innovation (RRI).

Overview of Responsible Research and Innovation

Scholars have analyzed the social dimensions of science and technology for centuries, and such analysis has become more formalized in recent decades. Key

[7] As an example of how non-anthropomorphic frames can be enlightening, see, e.g., Bryson (2015a) on why robots are more like novels than children.

disciplines that have contributed to this understanding include history, public policy, philosophy, economics, and especially science and technology studies (STS) (Felt et al., eds, 2017). Various events and trends in the 20th century contributed to greater attention to such issues. Revelations about Nazi medical experiments, for example, spurred calls for ethical treatment of human subjects; allegations of scientific misconduct in the United States created controversy in Congress and heralded greater oversight of federally funded research (Guston, 2000); and rising sensitivity to the military implications of science and innovation in the wake of Hiroshima and Nagasaki as well as the Vietnam War and other events (Moore, 2013) sparked greater debate in the scientific community about issues of social responsibility.

The United States government and others have sought to shape science and technology in various ways for centuries, but the US government has been especially explicit about this influence since World War II (Guston, 2000). More recently, there has been a substantial effort aimed at better anticipating and shaping innovation processes as well as outcomes, especially in Europe. While the United States government was for some time a pioneer of assessing the potential societal implications of emerging technologies, having an Office of Technology Assessment (OTA) for this purpose, OTA was terminated in the 1990s as part of broader government budget cuts (Bimber, 1996). In recent decades, European countries have taken the lead in technology assessment (TA), a precursor of RRI, and in public engagement with science and technology more generally. Countries that originally imitated the US's OTA, such as the Dutch and Danish governments, are now pioneers in methods for fostering democratic deliberation on the social impacts of technology. These countries have used various terms to refer to this work such as constructive TA (Schot and Rip, 1997).

Controversies such as the public debate over genetically modified organisms (GMOs), and many cases in which early signs of technology-related dangers were not heeded until substantial harm had occurred (Harremoës et al., eds., 2001), gave a strong impetus to calls for engagement with science and technology “upstream” in those technologies’ development (Wilsdon and Willis, 2004). Other events such as the Asilomar Conference on recombinant DNA also contributed to the heightening of responsibility discourse in the 20th century (cf. criticisms of the Asilomar model - Jasanoff, Hurlbut, and Saha, 2015). More generally, some have linked RRI in particular and the “responsibilization” of science and technology more generally (Dorbeck-Jung and Shelley-Egan, 2013) to the growing scale and temporal duration of technology’s potential impact, which calls for more sensitivity toward future generations and distant others than was required in earlier phases of human history (Jonas, 1979).

The theoretical and empirical literature on responsible research and innovation (RRI), and various associated practices, according to one account, stem from the synthesis of a number of other areas, especially STS, TA, and applied ethics (Grunwald, 2011). The term RRI and its recent antecedents or siblings such as “anticipatory governance” (Guston, 2014) stemmed in large part from rich discussion of the societal implications of nanotechnology and other emerging technologies in the 2000s. Other roots include the ELSI (ethical, legal, and social implications) and ELSA (ethical, legal, and social aspects) discourse stemming in the 1990s in the context of the Human Genome Project, which funded ELSI work in parallel with scientific work. When substantial U.S.

federal government investment in nanotechnology was being considered, Langdon Winner, an STS scholar, spoke before Congress about the need for a deeper integration between scientific progress and reflection on societal implications than had come before in the Human Genome Project (see Fisher [2005] on lessons from the Human Genome Project's ELSI program).

The legislation resulting from enthusiasm around nanotechnology ultimately funded two Centers for Nanotechnology in Society (CNS), one of which, headquartered at Arizona State University (CNS-ASU), pioneered a variety of conceptual and practical tools under headings such as anticipatory governance (Barben et al., 2008; Guston, 2014). CNS-ASU and affiliates have iterated the idea of anticipatory governance in other contexts such as energy (Davies and Selin, 2012), developed methods such as Socio-Technical Integration Research (STIR) involving the integration of social scientists and humanists into laboratories to encourage reflection on the social dimensions of research (Fisher et al., 2015). CNS-ASU participated in the popularization of the RRI framework (and associated concepts such as anticipatory governance) in the 2010s through the Journal of Responsible Innovation and the Virtual Institute of Responsible Innovation.

Rene von Schomberg, a policy entrepreneur in the European Union, played a key role in the emergence of RRI as a term, providing an early definition (von Schomberg, 2011). Von Schomberg led the European Union's integration of RRI into funding initiatives (Brundage and Guston, 2019). According to one interviewee quoted by Brundage and Guston,[8] von Schomberg was worried in the late 2000's that "the ethics of

[8] Note that Brundage and Guston (2019) was originally completed several years prior to publication in 2014. The associated IRB documentation for the project is included as an appendix, since the work was

research has become too internalistic, that it's just...FFP, fabrication, falsification, and plagiarism—and that that's all very internalist to within the scientific community. He wants to get the scientific community to be responsive to a social context in which they work…[R]esponsible research and innovation incorporates recognition and sensitivity to the social context."

The novelty (or lack thereof) of RRI compared to prior approaches to thinking about the social dimensions of science and technology has been characterized in various ways. Some have emphasized RRI's focus on the intent of research and innovation, rather than merely focusing on the products or processes of such research and innovation (Owen et al., 2013). Rip (2014) situates RRI in the context of an ongoing renegotiation of the moral division of labor in science (Douglas, 2009; Guston, 2000), and others such as interviewees of Brundage and Guston (2019) emphasize the comprehensiveness and coherence of RRI versus more piecemeal earlier approaches. Its advocates aim for the concept to be more comprehensive in its aims and prescriptions than its precursors, though they often caution that it is an evolving concept and should not yet crystallize or devolve into a mere checklist approach (Brundage and Guston, 2019). Scholars vary on what the novel aspects of responsible innovation consist of (Valdivia and Guston, 2015).

Stilgoe et al. (2013), in their influential account, define RRI as "taking care of the future through collective stewardship of science and innovation in the present." Those involved in the development of RRI concepts advocate a rethinking of the role of science and technology in society, and changes to the concrete practices involved in envisioning,

conducted under the auspices of Arizona State University. Other work reported on in this dissertation was conducted and registered as part of research projects hosted at other universities.

conducting, and disseminating related advances so that they are more beneficial, democratic, etc.[9] Like these scholars, I take as a foundational normative assumption that widely impactful technologies (or processes, more abstractly) should be governed in a way that is responsive to the interests and desires of those who are impacted. This explicit or implicit extension of the case for democratic influence over laws to democratic influence over technologies has been motivated in part by analyses of the law-like nature of technology in shaping human behavior (Winner, 1986; Jasanoff, 2016).

Frickel and Gross (2005) define a scientific-intellectual movement as a “collective effort to pursue research programs...in the face of resistance from others in the scientific or intellectual community.” Brundage and Guston (2019) argue that responsible research and innovation can be thought of as an instance of such a movement, as can more localized efforts within particular domains (such as efforts to foster greater responsibility in AI), in that they involve a group of scholars who perceive resistance among others toward the need to take the societal context of their work seriously, and engage in various forms of collective action to advance their cause such as writing letters to the editor, holding conferences, and establishing journals that focus on their area of concern. My work can be construed as part of this movement, advocating explicitly for more serious engagement with broader societal contexts and consequences of AI by researchers and other stakeholders.

RRI has informed initiatives in the EU (including the UK) and elsewhere, and now has a journal devoted to it (the Journal of Responsible Innovation). Recent

[9] Re: “democratic,” cf. Wong’s argument that responsible innovation should not be presumed to only apply in liberal democratic societies (2016).

discussion of the topic, much of it playing out in this journal, has extended and critiqued the concept and applied it to new domains. For example, authors have called for additional attention to the importance of care as a unifying concept for much of what RRI should be about (Grinbaum and Groves, 2013; Macnaghten et al., 2014; Kerr et al., 2017), and others have argued that RRI as a framework is limited in its ability to address salient features of the world of technology, such as "diverging and even contradicting interests" (de Hoop et al., 2016). One prominent framing of R(R)I, espoused by Stilgoe et al. (2013) in their paper, "Developing a framework for responsible innovation," highlights the dimensions of anticipation, inclusion, reflexivity, and responsiveness and the need for their integration (these are introduced in detail in the next chapter). This framing is not universally agreed upon, but in the remainder of the chapter and occasionally elsewhere in the dissertation, I draw on it since their work has attracted significant scholarly attention, with over 500 citations to date, and no competing list of dimensions has garnered widespread consensus. This framework, and RRI generally, was significantly influenced by earlier work on anticipatory governance (Guston, 2014). This connection can be seen in the fact that an influential work in the anticipatory governance literature (Barben et al., 2008) used a trichotomy that foreshadowed Stilgoe et al.'s four dimensions. Specifically, Barben et al. discuss foresight, engagement, and integration, where foresight is similar to anticipation and engagement is related to inclusion.

AI's Governance-Related Characteristics

There are several aspects of AI as a technology that are worth highlighting, since they bear on the tractability of its governance and what such governance might look like.

Furthermore, they suggest that RRI in AI will require a stronger emphasis on publication norms and coordination among actors than in other cases, and they highlight the urgency of developing improved means of anticipating AI futures as well as reflecting on unstated assumptions and path dependencies in the AI development process. I aim to partially fill this urgent need in the subsequent chapters.

Several governance-related characteristics have been previously highlighted in relevant literature, such as the capacity to evoke human social responses (Calo, 2015), AI's potential safety risks (Amodei and Olah et al., 2016), its privacy implications (Calo, 2011a; Brundage and Avin et al., 2018), its economic implications (Brynjolfsson and McAfee, 2014), and the risks of AI systems displacing human responsibility (Bryson, 2018). I pay particular attention here to the characteristic of AI's generality, as it is a major source of AI's appeal as well as its governance challenges. While many technologies pose safety, privacy, and economic risks, very few are general purpose technologies (Lipsey et al., 2005). General purpose technologies such as writing, electricity, and computers have distinctive societal implications (Lipsey et al., 2005) and are better positioned than ever before to spread rapidly. AI systems today can be diffused nearly instantly in an already globalized and Internet-connected world. I claim that the increasing generality[10] of AI systems (as defined below) makes publication norms a quintessential component of responsibility in the domain of AI. As a concrete illustration

[10] I use language here and elsewhere that suggests a *spectrum* of generality, with fully deterministic and single-purpose systems on one end and (physically impossible) fully general systems on the other. This is because, first, humans are not fully general, nor are any animals or artifacts--each is adapted to at least some extent to some set of tasks over others. Nevertheless, the effort required to adapt a human (or an AI system) to a new task varies, and e.g. speech recognition systems today are more easily and robustly adaptable than was the case several years ago. Second, there is no consensus on the right way to evaluate AI progress, as discussed later in the dissertation, and generality in particular is a contested concept.

of the confluence of generality, connectivity, and speed in AI, consider the speed with which published AI results are replicated, implemented at scale, and modified. The case of a specific AI system discussed later, GPT-2, illustrates these themes: upon the release of the largest and most performant version of GPT-2 in November 2019, it took less than one hour for the new version[11] to be incorporated into freely available online infrastructure, thus allowing anyone with an Internet connection to readily access and use to generate an extremely wide variety of (often human-passable) text.[12]

There are two senses in which AI might be characterized as general purpose: first, in the economic sense, even relatively limited aspects of intelligent behavior--such as learning to predict the right label for an image--are applicable across a massive range of societal contexts, as shown in the flowering of recent entrepreneurial, non-profit, and government-led efforts in this area in recent decades. Generality (in this sense of extremely diverse application domains) makes AI potentially enormously significant, particularly as it diffuses to and is tailored toward various particular industries and applications. Additionally, this form of generality suggests that AI governance will likely impinge on a number of other policy areas to the extent that it widely diffuses in society. A second sense of AI's generality involves not merely the applicability of a system to many societal contexts, but a system's ability to perform many functions "out of the box." This sense of generality is often linked to humans' ability to perform a range of tasks with less learning required per task: AI development today leans heavily on the use

[11] Smaller versions of the model had previously been released in a process known as staged release, discussed later.

[12] Specifically, the websites Talk to Transformer (www.talktotransformer.com) and Write with Transformer (transformer.huggingface.co) quickly upgraded to the new versions of GPT-2 within a day, and Talk to Transformer's upgrade specifically taking less than one hour.

of computing power to make up for the fact that machine learning is inefficient compared to humans. Systems with more “common sense,” it is sometimes claimed, would be more like humans in this sense of generality. For the purposes of most of the dissertation, it suffices to acknowledge that even contemporary (and earlier) levels of narrow AI capabilities are sufficiently powerful to have wide applicability and deep impact.

Language models[13] like GPT-2 are now capable of generating text that can deceive humans in many cases, just as image generation has recently matured to the point that visual deception is tractable. Generality in both senses is a spectrum, in the sense that no system (biological or non-biological) is capable of performing every computational task efficiently. Over time, due to a combination of improvements in algorithms, hardware, engineering infrastructure, and data, AI technology becomes more easily steered toward performing a more diverse set of tasks with less human intervention required for each additional task, compared to less general technologies (including in some cases earlier versions of the same AI system). For example, the language model GPT-2 is able to more efficiently adapt to new domains than earlier language models, and the larger versions of the system encapsulate more transferrable knowledge than the smaller versions (Radford and Wu et al, 2019; Solaiman et al., 2019; Brundage et al., 2019).

I focus on the first sense of the term “general” (i.e. on the wide range of applications for AI systems) as it is critical to understanding the contemporary landscape

[13] A language model is a system which predicts likely sequences of text based on observation of a large number of such sequences. Language models can be used to help analyze and/or generate natural language. GPT-2 (Radford and Wu et al., 2019) is an example of a language model.

of AI governance challenges, though I return later to the role of technical progress in shaping AI outcomes. And more generally, I will discuss the importance of understanding and navigating the range of different perspectives on AI's present and future.

General purpose technologies (GPTs) have been described by economists as the "engines of [economic] growth" (Bresnahan and Trajtenberg, 1992) - they lend themselves to substantial productivity-increasing applications and have impacts of substantial scale and duration compared to more limited technologies. Lipsey et al. (2005) consider various candidates for GPTs and identify only two dozen that meet the criteria laid out. These criteria relate to a technology having a fairly distinct technological core that transcends individual applications, a substantial scope for improvement, a variety of applications, and spillover effects. Electricity is a canonical example, and interestingly, AI is frequently referred to as "the new electricity" (see Brundage and Bryson 2016 for discussion).

Perhaps the most questionable criterion, as it relates to AI, is the distinct technological core, as AI research involves a range of methods, including tree search, reinforcement learning, clustering, etc. But AI defined broadly plausibly counts, and others have argued the case for a subset of AI (deep learning) counting as a potential future GPT. At a recent workshop organized by the National Bureau of Economic Research (NBER), several speakers argued for AI or a subset thereof counting as a GPT, and various implications of this framing were considered. Brynjolfsson et al. (2017) confidently claimed that AI is a GPT, substantiating this claim with reference to another set of criteria, from Bresnahan and Trajtenberg (1992). They argued that AI is, as that framework requires, pervasive, improvable, and able to spawn complementary innovations. They also noted that like other GPTs, AI might be expected to have a lagged economic impact as complementary innovations are discovered and AI-related capital is accumulated. Similarly, technologies like electricity took substantial time to be fully realized (and still, many lack access to it, a cautionary lesson for those trumpeting AI as the new electricity - Brundage and Bryson, 2016).

Brynjolfsson et al. went so far as to argue that AI is the most general of GPTs, in light of its ambition to replicate, augment, and surpass human cognition, which, in turn, is a key and pervasive input in the economy. Cockburn et al. (2017) explored the case for deep learning in particular as a GPT. In light of its fairly generic nature (mapping a range of inputs to a range of outputs through learning an assignment of weights in a neural network) and its wide applications, deep learning may count as a "new method for invention" according to Cockburn and colleagues. Finally, GPTs can interrelate in various ways, including via one enabling another. Digital computers are often classified as GPTs (Lipsey et al., 2005), and these enable AI: the value added by AI is in the creation of better software to run on such computers, and progress in computing hardware has been a major driver of recent breakthrough results in AI (Amodei and Hernandez, 2018).

AI's status as a GPT has several key implications for governance that have been under-emphasized in the RRI literature. While the RRI literature has in many cases grappled with potential or existing GPTs such as renewable energy or nanotechnology, the orientation of the literature has often been to critically assess the evidence for such generality being realized (e.g., Youtie et al., 2007) or to re-center the conversation on the limitations of the technology's potential contributions (Wiek et al., 2012). I instead start from the perspective that AI as a field is definitionally oriented toward building GPTs, and that there is already sufficient evidence to treat the field as building a cluster of general purpose technological capabilities, with enormous societal implications. While this framing still allows us to speak of degrees of generality and associated design choices, degrees and modalities of diffusion, and the role of human choice in designing technologies--all classic themes in the RRI account of technological governance--my GPT focus also facilitates a more direct reckoning with what makes AI an especially promising and challenging technology to govern today. And in particular, emphasizing AI's GPT characteristics is essential for grappling with emerging challenges related to publication norms (Crootof, 2019) and coordination in AI (Askell et al., 2019).

I turn now to the implications of AI's status as a GPT for what RRI means in this context, and begin to motivate the methodological approaches I take later in the dissertation, which is focused in particular on the RRI dimensions of anticipation and reflexivity.

First, a technology's status as a GPT is a prima facie reason to expect substantial societal implications across a variety of domains. Historically, GPTs such as electricity have been enormously impactful, contributing substantially to economic growth and influencing the distribution of income and wealth. Economic growth, in turn, has historically been associated with substantial changes in social mores (Friedman, 2005). Similarly, AI can be expected to increase economic productivity substantially, though with some lag (Brynjolfsson et al., 2017) as business models and processes are developed to productize and monetize AI. More generally, AI's impact on the future of work, education, and leisure (Brundage, 2015) raises a wide variety of governance questions.

Along with these pervasive impacts come pervasive governance questions, meriting a broad look at what tools, actors, and fora of governance are needed, as discussed further below.

Second, AI's generality suggests the inability to predict all possible positive, malicious, and ambiguous uses of AI, assuming that it continues to diffuse widely. As discussed later in the case of malicious uses, an expert elicitation exercise drew attention to a general class of concerns (generation of fake media) but did not anticipate all possible instances of this phenomenon, including some that occurred while the report was being written (fake pornographic videos with arbitrary faces inserted into them). While this is true of all technologies to some extent, GPTs are the extreme case of unpredictability, at least assuming a laissez-faire development process.

Note that inability to anticipate all possible applications or effects of a technology does not entail the technology's ungovernability: all technologies have this feature to some extent, and anticipation of possible, plausible, and preferred futures is distinct from prediction (Guston, 2014). However, it is a feature worth considering in the context of more encouraging governance-related features.

Third, AI's generality suggests that some applications can be anticipated in advance and deliberately pushed forward by private and public actors, on the assumption that a wide range of tasks can, given sufficient technical development, ultimately be automated. Thus, explicit efforts to increase (e.g.) mental health, manufacturing, or agricultural applications of AI are feasible. This raises the stakes of AI governance, and (contra the second point above), adds some foreseeability to AI's effects. It is possible for public or private actors to deliberately push AI progress forward in specific application areas, sometimes envisioned as technical fixes for societal problems (Sarewitz and Nelson, 2008).

Fourth, the underlying capabilities of AI are a critical dimension of possible futures with AI, thus suggesting a premium on efforts to anticipate plausible developments of underlying basic capabilities. For example, understanding progress in vision technologies is useful for understanding and shaping a range of possible applications, such as surveillance, medical imaging analysis, and a range of consumer services such as automatic tagging of photos in online photo albums. Anticipating in advance what sorts of technological applications are plausible would be highly valuable for anticipating the nature, severity, and sequencing of risks and opportunities, and shifting AI and its broader societal context in more positive directions. Independent of predicting or anticipating underlying capabilities, there is at the very least a high premium on surfacing unstated assumptions and disagreements about the underlying technical core of AI. As we will see later, this is a non-trivial challenge even if one focuses exclusively on technical dimensions of uncertainty. Views on the timing of future developments vary greatly. Such uncertainty poses challenges for policy analysts and practitioners in the field, for which a range of anticipatory tools should be explored. Recent work that I highlight later suggests that there are potentially significant gains to be had in terms of short-term forecasting of AI progress. While prediction is the focus of many in the AI community, as discussed later, prediction is distinct from the concepts of foresight in anticipatory governance (Barben et al., 2008) and anticipation in RRI (Stilgoe et al., 2013).

Fifth, AI’s generality lends itself to various malicious applications as well as beneficial ones, making it a dual-use technology (Tucker, ed. 2012; Brundage and Avin et al., 2018), with various governance implications. As a GPT, AI will likely be used pervasively by criminals, terrorists, militaries, and intelligence organizations, raising challenging questions of conflict, power, and responsibility. Among other implications, the dual-use nature of AI gives rise to important questions related to the security implications of publishing AI research (Brundage and Avin et al., 2018; Dafoe, 2018; Solaiman et al., 2019) and the need for cooperation among AI developers (Askell et al., 2019). In the next section, I will discuss some such implications when discussing the release of the language model called GPT-2 by OpenAI.

Beyond generality, which looms large in my framing of RRI for AI, two additional governance-related characteristics of AI are worth highlighting.

A first additional governance-related characteristic of AI is the very unequal concentration of key inputs today. Talent is widely considered to be scarce (Kahn, 2018; Gagne, 2018), and computing power is unevenly distributed, with experiments by organizations such as Google Brain and DeepMind routinely using hundreds or thousands of computing cores, compared to single or double digits elsewhere. Access to insider tacit knowledge about engineering best practices is unevenly distributed and only sometimes described in publications. Access to relevant data is unevenly distributed. Funds to purchase the above inputs is, of course, also unevenly distributed (Piketty [2014] notes both the high concentration of capital and its tendency to become more concentrated over time). AI research is typically very open, but it is unclear whether this will or should last forever (Bostrom, 2017; Brundage and Avin et al., 2018), as the field grows and grapples with the potential for AI misuse.

Second, AI, as a subset of digital technology, is highly scalable (Brundage, 2018b): copies of AI systems can be produced at much lower costs than their design and training costs, those copied systems can often be run at very high speeds, and the throughput of large-scale systems, e.g., for image recognition on Facebook, or machine translation, is often very high. Additionally, AI is increasingly being deployed on ubiquitous devices such as smartphones, enabling this digital scalability to translate into physical scalability - AI can quickly be deployed to billions of users worldwide. This scalability raises additional issues related to concentration of power, since sufficiently capable AI enables the straightforward conversion of capital into labor and thereby the concentration of economic productivity in a small number of hands. It may also enable large-scale drone swarms, automated hacking, and digital surveillance (see Brundage and Avin et al., 2018 for a review of such concerns). Finally, scalability can be used to produce widespread economic growth and prosperity (Brundage, 2018a), but this is not a foregone conclusion (Sachs et al., 2016).

In summary, AI is unevenly distributed, scalable, and general purpose. The ability to anticipate its development is desirable but also potentially quite challenging given expert disagreement (Grace et al., 2017), an incomplete theory of AI's progress (Brundage, 2016a, Hernandez-Orallo, 2016), and exogenous factors like hardware developments (Hwang, 2018).[14] AI's generality and scalability enables its deliberate application to a wide range of positive purposes, but also makes forecasting its malicious uses and other forms of deleterious effects highly difficult.

The Importance of Publication Norms in RRI for AI

The above characteristics suggest the strong need for responsible publication norms in the field of AI, as well as the potential challenges for coordination among different actors to govern AI. The fact that AI practitioners routinely design, deploy, and/or publish systems that can be used for a range of purposes is exciting from the perspective of those eager to build/use beneficial systems. But it can be just as exciting for those eager to misuse technology in the pursuit of profit, political power, or ideology, as discussed in recent debates surrounding misuse of language models (Solaiman et al., 2019). The GPT nature of AI implies a need for reckoning with publication, but norms have not yet caught up with the growing societal impacts of AI. Specifically, there is a strong norm of openness in the AI community, and recent events have caused some of these assumptions to be questioned in light of AI systems' growing malicious use potential.

Note that by centering publication norms in RRI for AI, I am not implying that the deployment of specific AI systems in commercial or other contexts loses its importance. Rather, the extreme diversity of possible AI applications that could emerge from a given base system--including generation of text for poetry, programming, and prose in the case

[14] The current deep learning "boom" in AI emerged partly as a result of the repurposing of existing computational resources to AI - specifically, GPUs, or graphics processing units, which were originally developed for videogames.

of GPT-2 (Solaiman et al., 2019)--is a consideration that the AI community must grapple with alongside other concerns.

Finally, note that I am not claiming that publication norms are only relevant in the context of AI--vibrant debates on openness have occurred in biotechnology and cybersecurity, for example. My claim is rather that the nature of AI, combined with additional facts about the world today such as the global nature of the AI ecosystem and the pace of iteration and repurposing of systems, demand substantial attention to the ethics of publication and the challenges of cooperating on responsible AI development.

To deepen the case for integrating insights from RRI with the particular governance challenges of AI development, I briefly recapitulate some recent developments in AI publication norms. These developments illustrate the kind of challenges AI poses as a general purpose technology and help illustrate the urgency of methods discussed later such as scenario planning, expert elicitation, and formal modeling.

In 2018, I co-authored a report calling greater attention to the potential for malicious uses of AI (Brundage and Avin et al., 2018). The report highlighted that the same technical systems would in many cases be applicable to both malicious and beneficial purposes. At the time, “deep fakes” were beginning to emerge in the public consciousness (Cole, 2018a) and our report highlighted this and other instances in which synthetic media could be used to deceive people. The potential for AI to scale up such malicious acts beyond what was possible previously was the subject of some further scholarly discussion, though it did not apparently influence publication norms or other behavior directly.

Norms began to change at a faster pace the following year in response to a concrete case: the GPT-2[15] language model was announced by OpenAI, and the debate it sparked highlighted the urgency of grappling with AI's dual-use nature (Crootof, 2019). Ongoing efforts by OpenAI, other actors in the language model space discussed above, and the Partnership on AI aim to foster an AI community-wide discussion on publication norms (Partnership on AI, 2019), using some aspects of the GPT-2 case as inspiration. The central dilemma of the GPT-2 case is that the very same language models can be used for creative, commercial, scientific, or malicious purposes. In OpenAI's initial communications on this topic, they highlighted the inherent generality of language modeling as a sub-field of AI (Radford and Wu et al., 2019): language models trained on large fractions of the Internet naturally acquire a range of capabilities that afford commercial and creative applications, since their objective is simply to predict what comes next in strings of text. Given the diversity of text on the data and the growing capacity of large language models to capture the richness of this data distribution, language models are increasingly able to generate synthetic text that is perceived as credible to humans. Research has shown that there is significant variation even among different sizes of the GPT-2 system, suggesting strong potential for further improvement (Solaiman et al., 2019), as one would expect from a general purpose technology. In particular, the GPT-2 paper (Radford and Wu et al., 2019) demonstrated fairly consistent

[15] GPT here standards for Generative Pre-trained Transformer - no relation to the GPT concept in social science.

returns to scale, suggesting the possibility of further improvements in performance from scaling up to more data and computing power.

Subsequent to the initial announcement of GPT-2, which sparked significant discussion in the AI community (Partnership on AI, 2019), other language models were subsequently published by other laboratories. Examples of subsequent models released following GPT-2 include GROVER (Allen Institute for AI and the University of Washington), and CTRL (by Salesforce). Some aspects of these subsequent developments mirrored OpenAI's approach, such as the lengthy defense of one's publication approach (vs. defaulting to a presumption in favor of publication) and in the case of GROVER, a "staged release" approach for sharing the model as well as providing earlier access to researchers than to the general public. Further information about this case can be found in Crootof (2019) and Solaiman et al. (2019). I return to the GPT-2 case later in order to illustrate the concrete and pressing challenges involved in RRI for AI today, and to contextualize some of my recent contributions.

The Importance of Cooperation in RRI for AI

Beyond highlighting the importance of publication norms, AI's status as a general purpose technology also highlights the potential importance of coordination among AI developers. Since AI can be put to a very wide range of purposes, scrutinizing the way in which generic computing power and technical skills are used will be of increasing importance to AI governance. In the context of international relations, for example, investments by governments in AI capabilities can be threatening to other countries who fear militarization of those capabilities. Finding ways to credibly signal one's intentions in the context of AI development, and developing appropriate accountability procedures to ensure that capabilities are not being put toward dangerous uses, is thus essential (Bryson, 2019). While a need for coordination is not a novel aspect of AI--the importance of collective action has previously been highlighted in the RRI literature at an individual level (Spruit et al., 2016) and is implicit in RRI accounts that emphasize multi-stakeholder governance--the game-theoretic dynamics of technological development have not been extensively explored in RRI to date. I begin to address this gap later in the dissertation. In particular, I describe preliminary efforts in the direction of formally modeling cooperation problems in AI development, which have subsequently been further developed in the literature (Askell et al., 2019).

Finally, note that the need for publication norms and the need for collective action in AI are not unrelated. In the case of GPT-2's publication, coordination across AI developers on the release of similar systems was essential (Solaiman et al., 2019). OpenAI and other organizations such as the Allen Institute for Artificial Intelligence, the University of Washington, Facebook, Salesforce, and others exchanged perspectives on language model properties and implications, and OpenAI began advocating a norm of prior notice before publication of AI systems in similar contexts (Solaiman et al., 2019). There are individual and organizational incentives to release AI systems, such as garnering recognition, but uncoordinated publication will likely cause negative externalities (Brundage and Avin et al., 2018), such as rampant misuse of language models for disinformation with inadequate attention to detection and public education. In later chapters, we return to the challenges of responsible publication and cooperation in

AI by discussing my efforts to better anticipate and reflect on such issues in an increasingly disciplined way.

Conclusion

AI has a number of governance-related characteristics that demand attention from anyone seeking to understand the challenges and affordances of its governance, or to take concrete steps to govern it. Chief among these is the general purpose nature of the technology, though others such as the highly concentrated nature of key inputs and the scalability of AI are also important. Responsible research and innovation (RRI), a broad framework for reasoning about scientific and technological governance, serves as the foundation for my analysis. Existing concepts and methods in the RRI literature need to be augmented with a greater focus on generality in order to fully grapple with AI, and this perspective leads to a focus on publication and coordination as critical challenges for responsibility. Using the existing RRI framework to evaluate the state of AI governance is the task of the next chapter. Subsequently, I turn to my practical efforts to expand capacities for anticipation and reflection in AI, given some of the governance challenges and opportunities discussed above.

CHAPTER 3: AN ASSESSMENT OF EXISTING EFFORTS

Introduction

In this chapter, I critically review the recent history of AI governance through the lens of RRI in order to identify gaps to be filled via a more disciplined approach to AI governance. I begin by providing an overview of AI governance, and then separately discuss extant work through the lenses of anticipation, inclusion, reflexivity, and responsiveness. In each case I compare the state of AI governance today to the more ambitious aspirations in the RRI literature. The following chapter moves into the realm of RRI methodology and asks what methods we might use to better anticipate diverse plausible AI futures and to embed more reflection in the AI development process.

Overview of AI Governance

Informal discussions of the appropriate governance of AI are almost as old as the field itself, though little systematic scholarly attention was paid to it until the late 20th century. Pioneers such as Wiener (1964) and Weizenbaum (1976) wrote critical early warnings about the societal implications of their fields, and in arguing for AI's potential. Alan Turing also was attentive to such broader implications. Anticipating future societal challenges, Turing rebutted various objections to the possibility of AI, noted the critical importance of human input even in the context of learning machines (at a time when machine learning was not yet a proper field), and anticipated a change in the use of terminology over time as machines become more capable (Turing, 1947).

Since these early days, there has been continued discussions of AI governance issues in journals such as Artificial Intelligence and Law (published since 1992) and AI & Society (published since 1987). Artificial Intelligence and Law was described at its launch as “devoted to artificial intelligence and law, an interdisciplinary field that combines one of the oldest human intellectual endeavors with one of the youngest” (Editors of Artificial Intelligence and the Law, 1992[16]). AI & Society “[provides] a forum for exploring the likely effects of these new technologies, and for debating policy and management issues” (Editors of AI & Society, 1987[17]). While each of these has played some role in the fostering of discussions on AI governance, and I cite some work published in AI & Society below, neither achieved a strong and long-standing place in such discussions.[18] For example, legal issues related to AI have more recently been vibrantly explored at the We Robot conference series, which publishes its own articles.

Additionally, there have occasionally been keynote talks and workshops at major AI conferences such as Association for the Advancement of AI (AAAI), European Conference on Artificial Intelligence (ECAI), and International Joint Conference on Artificial Intelligence (IJCAI) on societal impact-related topics, with the frequency of such speakers and discussions increasing in recent years. For example, AISB 2000 held a “Symposium on Artificial Intelligence, Ethics, and (Quasi-)Human Rights (Barnden et al., 2000) and in recent years, AAAI has had an annual workshop on “AI, Ethics, and

[16] Interestingly, the identities of the original editors, and the authors of the first introductory article, could not be easily found online via the journal’s website.

[17] See footnote above.

[18] While I have investigated the fate of these efforts in detail, it seems plausible that momentum stalled due to a limited perception of urgency. Overall public and private attention to AI, including attention to its societal implications, has significantly increased in recent years, although further work would be required to account for, e.g., *AI & Society*’s fate in detail.

Society", which more recently evolved into a standalone conference. Sub-fields of AI like machine learning have also seen increases in such activity in recent years, such as the International Conference on Machine Learning (ICML) and the Conference on Neural Information Processing Systems (NeurIPS). Machine learning conferences have significantly increased in attendance in recent years, concurrently with the spreading of the "deep learning revolution." As a result, there are frequent symposia, workshops, and tutorials on governance issues at contemporary machine learning conferences.

Overall general public attention to AI has increased significantly over the past decade, as documented by Fast and Horvitz (2016) using data from the New York Times's coverage. In particular, they find a substantial increase in coverage starting around 2010, with different foci during previous eras. For example, chess was more widely discussed in the Deep Blue era of the 90's, and "doomsday" appears as a keyword in the last decade. The past five to ten years has seen the rise of deep learning as a much-hyped subset of AI. AI has attracted enormous amounts of scholarly, government, and commercial interest, and alongside this growth has come a significant growth in discussions of ethics, responsibility, and policy. Even before the current boom of excitement, there were early steps in the direction of more rigorous reflection on AI governance, such as the development of the EPSRC principles, the work of a team of researchers chartered by a leading UK funding council (Boden et al., 2011; Bryson, 2017). Still, interviewees quoted by Brundage and Guston (2019) agreed that progress in AI helps explain the recent uptick of interest in AI's societal dimensions.

Several trends related to the discourse on AI governance in the past five years should be highlighted, during a period that Tom Dietterich, then the President of AAAI, and Eric Horvitz, lead AI research manager at Microsoft, described as the "rise of concerns about AI" (Dietterich and Horvitz, 2015).[19]

First, the AI community and adjacent fields have increased their attention to issues of fairness, accountability, and transparency, sometimes referred to as "FAT" issues. The FAT community has hosted a regular conference in the past few years called FATML (and more recently, FAT*) that has seen spiking attendance, and numerous papers have been published on issues such as machine learning systems' learning biases from human-generated data (Caliskan, Bryson, and Narayanan, 2017), the fairness of algorithms' decision-making (Hardt et al., 2016), and the transparency (Weller, 2017) and accountability (Kroll et al., 2016) of machine learning systems and software more generally. Such issues have also been discussed in mainstream media (e.g., a Google system's misclassification of African Americans as "gorillas" received widespread news coverage), influencing and being influenced by the scholarly attention to FAT issues.

Second, there has been increased attention to AI safety in various senses. This attention has been distributed between the present and near-term safety risks associated with driverless cars, drones, autonomous weapons, as well as more speculative concerns falling under labels such as "the control problem" or "the value alignment problem" (Bostrom, 2014). The control problem was discussed in Bostrom's 2014 book Superintelligence, which made the case for the difficulty and importance of aligning AI systems' behavior with human values, a message echoed in more technical detail in a

[19] Given that there is a rich prior history of ethical reflection in and about the field of AI, this framing may suggest more novelty in recent discussions than there is, but "rise" is certainly an accurate reflection of the direction of change.

paper entitled “Concrete Problems in AI Safety” by researchers at OpenAI, Google Brain, and Stanford (Amodei and Olah et al., 2016). Such concerns have been discussed widely in the media (Fast and Horvitz, 2016), though they have also been criticized by some researchers as being misguided for various reasons, such as being preoccupied with risks that are excessively far in the future, misframing the nature of AI, being impossible to study fruitfully today, or distracting from more pressing problems (Bryson, 2013; Lawrence, 2015; Etzioni, 2016; Ng, 2015; Crawford and Calo, 2016).

Safety concerns have a long history in AI. For example, Weld and Etzioni in 1994 (20 years before Superintelligence) wrote that “society will reject autonomous agents unless we have some credible means of making them safe.” The research area of reinforcement learning considered problems such as safe exploration, and control theorists designing complex systems have long been concerned with safety. Even further back, Turing, Wiener, and other scientific pioneers also recognized the deep ethical issues posed by AI systems. Wiener, for example, foreshadowed contemporary concerns about AI safety: “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” (Wiener, 1960). Note that the presupposition of a natural trajectory toward AI systems “with whose operation we cannot interfere effectively” is also controversial (Bryson, 2019).[20] Finally, note that recent discussions and actions on the bias and transparency of AI systems are closely

[20] Bryson argues, for example, that through careful system engineering with clear specification of capabilities, tasks, inputs, and outputs, AI developers can trivially avoid many of the failure scenarios envisioned by, e.g., Bostrom.

related to safety, in that they involve building assurance that systems will act as intended in a range of situations.

The relationship between contemporary safety risks from existing systems and longer-term, more speculative concerns has been discussed by various authors--some have noted substantive connections between the research questions involved (Krakovna, 2018), while others have noted the convergence of interests between those advocating for safer systems (Baum, 2017). Authors such as Geraci (2012) and Finn (2017) have mapped some of the broader cultural influences and consequences of such debates.

Third, there has been a substantial increase in attention to AI governance from national governments (e.g., the US, the UK, Canada, China, and Japan, among others), subnational governments (e.g., California and other US states re: driverless cars, Chinese and Canadian cities and provinces re: fostering AI research and commercialization, and New York re: algorithmic accountability), and policy-oriented transnational organizations (e.g., the Organization for Economic Co-operation and Development [OECD], the Institute for Electrical and Electronics Engineers [IEEE], and the World Economic Forum). Notable outputs of this work include the “Preparing for the Future of Artificial Intelligence” report from the White House in 2016 (Executive Office of the President, 2016), a series of reports and events from OECD and the World Economic Forum, the IEEE’s Ethically Aligned Design series of reports, the Pan-Canadian AI Strategy, and substantial increases in government investment and activity in China (Ding, 2018). The OECD principles in particular have been prominently signed on to by the United States and other AI-relevant countries.

Finally, subsets of the AI community and academics in other disciplines have engaged in various forms of collective action aimed at raising awareness about these issues among the community itself and society more broadly (Brundage and Guston, 2019). Collective action efforts, of which the EPSRC principles mentioned above are also an early example, include the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (drawing on contributions from hundreds of researchers[21]; IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2018); the development and highlighting of the aforementioned AI safety research agendas (Amodei and Olah et al., 2016); and the FATML conference series and associated work such as bibliography creation (FATML, 2018). As another example, the Future of Life Institute (founded by a mix of entrepreneurs, physicists, AI researchers, and others) sponsored conferences in Puerto Rico and Asilomar and organizing several open letters on AI, each deliberately launched to media fanfare. The first, "Research Priorities for Robust and Beneficial Artificial Intelligence" (Future of Life Institute, 2015), highlighted AI safety as well as the distributional impacts of AI, and called for research, subsequently funded after an RFP, on related issues. The Partnership on AI to Benefit People and Society is a deliberately industry-spanning and cross-sectoral example of collective action: it brings together a wide range of industry and civil society organizations to "study and formulate best practices on AI technologies, to advance the public's understanding of AI, and to serve as an open platform for discussion and engagement about AI and its influences on people and society" (Partnership on AI, 2018). Finally, at an international level, the

[21] I have participated in this initiative as a member of the IEEE Law Committee.

adoption of OECD principles on AI was a major step forward with respect to collective action.

While much of the RRI discourse and the AI governance discourses have proceeded mostly independently (Brundage and Guston, 2019), one exception is my (2015) proposal of a three-fold conception of responsible innovation in AI, which suggests that researchers and engineers ask and answer the following questions:

- **First**, how could different kind(s) of AI affect society, and how should this affect research goals in the present?
- **Second**, what domains should AI technology be applied to/how urgently, and what mix of basic and applied research is optimal from a societal perspective?
- **Third**, what does the public want from AI, and what do they and various stakeholders such as policy-makers, educators, [parents, students, businesspeople, etc.] need to know?"

This framework has informed some of my subsequent research, e.g., my writing on specific domains for AI applications (Brundage and Danaher, 2017) and my concern with public engagement (Brundage, 2016b). The questions map on roughly to the RRI dimensions of anticipation/reflexivity, reflexivity/responsiveness, and inclusion, respectively, discussed further below. However, note that these questions do not specifically address publication norms, and questions such as "what domains should AI technology be applied to/how urgently" only give a weak sense of AI's GPT nature. As a result, I would now add a fourth question to this list: How can malicious or otherwise harmful uses of AI be prevented or mitigated, and how can AI developers and other actors cooperate to ensure that such systems are published and deployed responsibly?"

Anticipation

In this and the following four sections, I review work done on AI governance falling under the headings of anticipation, inclusion, reflexivity, and responsiveness, and in each case point out the limitations of existing work. The limitations of extant approaches demonstrate the potential opportunity from taking RRI more seriously as a guiding framework for AI governance.

Anticipation, in the context of RRI, refers to the deliberate exploration of diverse futures for an area of science and technology, including both technical and societal elements. The perceived need for such thinking is longstanding, both in academia and more broadly, with organizations sometimes being designed to perform such tasks in an official capacity for governments (Sadowski, 2015). More recently, researchers have developed the concept of anticipatory governance, "a broad-based capacity extended through society that can act on a variety of inputs to manage emerging knowledge-based technologies while such management is still possible" (Guston, 2014), and applied it to fields such as nanotechnology (Barben et al., 2008; Davies and Selin, 2012) and synthetic biology (Brian, 2015). As described by Barben et al. (2008), anticipatory governance comprises multiple dimensions, with "foresight" being closely related to what's discussed here. As distinct from forecasting, foresight (and anticipation, as I use the term here) seeks to explore multiple plausible futures, with the aim of informing actions to make preferable futures more likely, while eschewing single point predictions.

There is a growing body of work on possible AI futures in the form of roadmaps, analyses, or predictions (e.g., Armstrong et al. 2016; Christensen et al., 2016) of particular futures considered more or less likely. However, notably, there is little methodological similarity between these various efforts, and little effort to integrate different conceptions of the future into a clear decision-making framework. I discuss in the next chapter how this state compares to anticipation in the energy sector in order to further illustrate the need for improved anticipation in AI.

Analyses have detailed various factors and dimensions of AI progress, with an emphasis on the diversity of possible ways that it could develop (Bostrom, 2014; Brundage, 2016a; Yudkowsky, 2013). For example, much has been written regarding the possibility of an intelligence explosion (Good, 1964) or "fast takeoff" (Bostrom, 2014), in which AI develops quickly from a stage in which it is in the same general vicinity of human intelligence in various dimensions, to significantly beyond human levels (cf. Bryson, 2013). Others such as Brooks (2014), Hanson (Hanson and Yudkowsky, 2013; Hanson, 2016), Ng (2016), Bryson (2013, 2019), and Dietterich and Horvitz (2015) have expressed skepticism regarding such concerns. The pace of development in AI capabilities is one dimension along which AI futures may vary, but it is not the only one. The right way to characterize this pace of development is highly disputed.[22]

Concerns about the future of AI have ranged from what Amodei et al. (2016) call "accident" risks, resulting from mistakes in AI design, to the intentional creation of destructive AI systems (Brundage and Avin et al., 2018), to more subtle risks. A broad range of scenarios and preferences have been expressed about AI futures, though as yet

[22] For example, Grace et al. (2017) find a range of perspectives on future developments, and Grace et al. (2016) find that respondents also vary in the extent to which they credit recent advances to data, algorithms, or hardware advances.

there is no agreed upon means of analyzing them or identifying solutions across a range of possible scenarios, a gap I discuss further in the remaining chapters. There has also been a fairly limited scope of such analyses, focusing largely on technical trends and extreme outcomes, with less attention to more complex, second-order effects of an increasingly automated society (though cf. Bryson, 2015b; Danaher, 2017). In addition to statements of particular views, there have also been surveys focused on experts' expectations of AI outcomes (Grace et al., 2017), though, as discussed further later, there has been no rigorous exploration of the range of views on experts' preferences for managing those outcomes, a gap I seek to begin rectifying in the chapter on expert elicitation. Finally, note that a variety of possible futures related to AI have been explored in science fiction, and these are often referenced as inspirational by researchers in the field (e.g., Data from Star Trek, R2-D2 from Star Wars, or various robots from Asimov's fictional universes).[23]

Inclusion

Inclusion, in the context of RRI, refers to the incorporation of the views, preferences, and interests of a wide range of stakeholders in decision-making about research and innovation. The emphasis on inclusion in technological governance has longstanding roots and many concrete forms such as lay participation in decision-making panels, citizen juries, and deliberative polling (Stilgoe et al., 2013). Inclusion is closely related to the aspiration of deliberation about possible futures, and indeed early

[23] For example, Cynthia Breazeal has repeatedly noted the influence of R2-D2 and C3PO on her career (Breazeal, 2011).

articulations of frameworks for RRI used this term rather than inclusion (Owen et al., 2013). This reflects the influence of ideals of deliberative democracy on thinkers in this area, as further evidenced in some of the concrete cases below.

There have been some limited efforts at inclusion in the case of AI, but generally, most discussions of AI governance have not taken seriously the long-standing critiques of purely symbolic inclusion exercises (Stirling, 2008). There have been some limited efforts toward making AI innovation more inclusive, which I discuss below. To begin, I'll contrast inclusion in AI with a reference case, the NASA-ECAST events engaging the American public on conversations regarding asteroid risk mitigation and asteroid exploitation. This initiative, in which Arizona State University participated[24], featured two in-person events and an online discussion, in which a representative sample of lay people were brought together to learn fundamental information about asteroids, explore scenarios for possible futures for NASA's and other parties' role viz-a-viz asteroids, and ask questions of experts (Tomblin et al., 2017). This exercise was motivated by NASA's internal decision-making process, in which they had multiple plausible options to pick from. Dimensions along which this exercise, though imperfect, was well designed, include: thoughtful preparation of introductory materials, debiasing of participants' responses to expert input via anonymization and text-formatting of responses to queries, active facilitation of discussions, and decision-relevance. In contrast, a notable recent case of AI inclusion, the White House's series of AI workshops in 2016 which were ostensibly aimed in part at engaging the public, two of which I participated in, was less

[24] I played a minor role as a facilitator of discussions at the Phoenix event.

well developed along these dimensions (Brundage, 2016b). In particular, they featured a fairly limited range of opinions within events on the landscape of possible futures, were largely lecture- and Q&A-based rather than more deeply interactive, and the presentation of content to non-experts seemed more based on the idiosyncrasies of individual presenters rather than deep reflection on the necessary knowledge and caveats that participants should be aware of, as was clear in the NASA case. Other public engagement exercises have been carried out in AI, with varying degrees of formality, publicity, and success, and (non-interactive) public surveys are commonly carried out by EU Barometer and others. Recent steps in the direction of more serious engagement in AI include the "Our Driverless Futures" effort (Farooque and Kaplan, 2019) in the US and other countries and an effort by the Royal Society for the Arts (RSA) and DeepMind in the UK (Balaram et al., 2018).

Finally, inclusion in the case of AI has often taken on a particular connotation, namely improving the gender and racial diversity of the field itself, which has been characterized as having a "sea of [white] dudes" problem (Clark, 2016). While the AI community has made some progress in at least acknowledging this problem in recent years, it's not yet clear what if any impact there will be of this acknowledgment. Several initiatives such as AI4ALL (AI4ALL, 2018), Women in Machine Learning (WiML, 2018), and Black in AI (Black in AI, 2018) have been launched to directly target these issues, and have likely positively impacted many individuals,[25] but much more will need

[25] For example, anecdotally, I have heard many accounts of the Women in Machine Learning events being successful in increasing perceptions of inclusion, and of having generally high quality events. Such events feature female and non-female speakers, as well as mentoring opportunities, at major machine learning conferences.

to be done. Critics of inequality in the computer science community have long been aware of systemic problems in the field (Hicks, 2017), decades before the current wave of awareness, and recent public controversy about sexual harassment at machine learning conferences suggests there is still a long way to go (Bergen and Kahn, 2017). Further, demographic representation in a technical field is just one aspect of the inclusiveness of a technical field, and further effort is needed to attain direct input of affected communities into design decisions. As noted by a growing number of scholars, including those in the FATML community, the harms associated with AI often fall disproportionately on those lack various forms of privilege (Crawford and Calo, 2016; O'Neil, 2016). Examples of such disparate impacts include biases in search engines, image recognition, automated CV screening, and predictive policing (O'Neil, 2016; Ferguson, 2017). Finally, while there is a growing trend of scholars devoting their attention to "AI for good," there has been no systematic effort to map out precisely how AI can be leveraged most effectively for good--nothing like a "GiveWell for AI"[26] exists--and the discourse surrounding AI for good is often muddled (Malliarki, 2019; cf. Floridi et al., 2018).

Reflexivity

Reflexivity, in the context of RRI, refers to awareness of the assumptions one is making about the science and technology one is working on, including its societal implications. This self-awareness at individual and group levels is critical to modulating those assumptions in response to evidence, and ultimately (discussed further below in the

[26] GiveWell is an organization that rigorously analyzes the effects of charities with an eye toward informing donors.

subsection on Responsiveness) acting on assumptions that are justified. Alas, it is common for researchers to take certain assumptions for granted, e.g., those that are “inherited” by a certain institutional culture or advisor-advisee relationship, and to be unaware of the diversity of plausible opinions. As Fisher et al. (2006) argue in an influential account of reflexivity, the “midstream” of innovation is rife with these kinds of decision points and opportunities for greater reflexivity. Inculcating awareness that scientists and engineers are in fact making normatively significant decisions in their work, and equipping them to act responsibly on this awareness, are critical challenges.

There has been little discussion of reflexivity in the literature on AI governance, but some findings in the survey by Grace et al. (2017) shed light on this. The extended results of the survey (published as Grace et al., 2016) highlight the diversity of expectations held by AI researchers, and their lack of reflexivity about this diversity. Specifically, AI researchers tend to incorrectly believe that others share their views about how long certain developments will take, when in fact they differ markedly. This can perhaps be explained by some combination of the limited explicit discussion of underlying assumptions, and the lack of sustained attention to long-term and/or societal issues by many researchers “in the trenches” focusing primarily on their next contribution.

Responsiveness

Additionally, there is the question of AI researchers (and others in the ecosystem of AI research and development) actually responding to their anticipations, reflections, and inclusion of a wide range of participants. The actual adherence of researchers to codes of conduct is a longstanding issue in the ethics of science and technology, and debated in contexts such as training of engineers, but responsiveness is broader than this. In concrete terms, responsiveness of a technical community to societal dimensions of their work requires more than merely reflecting on these dimensions at the proverbial pub at the end of the conference, but doing things differently as a result.

In this respect, AI governance has been highly limited. For example, like the RRI framework discussed above (Stilgoe et al., 2013; Brundage, 2015), AI governance prescriptions have typically been very high level and disconnected from concrete day-to-day decisions. The much-touted Asilomar Principles, developed at one of the aforementioned events organized by the Future of Life Institute, while they may be valuable, are fairly abstract, such as: "The goal of AI research should be to create not undirected intelligence, but beneficial intelligence" (Future of Life Institute, 2017). And while the FATML community has developed many rich accounts of algorithmic tradeoffs between different conceptions of fairness, for example, there is little in the way of established norms for implementing such techniques in practice.

Conversely, while some governance principles are fairly specific, they have not seen universal adoption. For example, the EPSRC principles prescribe that robots "should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent." And yet, as regular news stories attest, some entrepreneurs flout this norm with what are arguably public relations stunts. For example, "Sophia" by Hanson Robotics is an animatronic robot that masquerades as an entity with interests of its own, and was given Saudi honorary "citizenship." Some more recently have called for renewed attention to this concern; e.g., Tim Wu (2017) called for a "Blade Runner Law," essentially a legalization of a 2011 EPSRC principle. No such law has been passed anywhere to my knowledge.

Finally, while there has been an uptick in attention, and in many cases scholarly and industry time devoted to, societal issues, it is unclear whether much has changed in the concrete practices of researchers or the shaping of societal outcomes. There is undoubtedly more work happening on these topics in absolute terms, but whether renewed attention and advocacy has substantially changed the extent to which individual researchers, on average, engage with such issues is unclear. Simultaneous to these developments, the amount of effort devoted to developing AI capabilities has also increased substantially, as indicated by various metrics curated by the AI Index project (Shoham et al., 2017). As an illustrative example of limited responsiveness, several researchers whom I've interviewed (Brundage and Guston, 2019) have informed me that recruiting researchers to work on AI safety remains highly difficult, and that improving AI capabilities remains "sexier" in the eyes of much of the community.

Finally, a marker of limited responsiveness in the AI community is the way that the "politics of novelty" has played out (Guston, 2014). As has been noted by Guston and others (e.g., Stilgoe, 2016), the novelty or lack thereof of a technology can be characterized in a way that is not necessarily logically defensible, but which has the result of fending off any challenges to the field's autonomy and self-governance. A technology may be characterized as novel for the purpose of garnering funding, but mundane when issues of governance arise. Brundage and Bryson (2016) note that some have said that it is too early to have AI governance, e.g., because too little is known or because the technology is immature--despite AI already having substantial societal effects, and there already being substantial de facto governance. Others have stated that claims related to the ethics of AI could also be said about the ethics of databases (Evans, 2018)--a stance that both elides the ways in which AI is more potent than earlier information and communication technologies (ICTs) by virtue of its better handling of uncertain inputs and greater autonomy, and glosses over the ongoing harms associated with poorly governed "dumb" ICTs as if they are no longer a live issue. However, some have taken this "politics of novelty" in what are arguably more productive directions, emphasizing the lack of need for fundamental breakthroughs beyond existing technologies in order for AI to be misused, and the continuity of present and future systems (e.g., Brundage and Avin et al., 2018).

Conclusion

Examining the recent history of AI governance through the lens of RRI is illuminating and points to areas for improvement. In particular, we can see the paucity of rigorous comparison of different anticipated futures; the limited spread of reflexivity; the remaining limitations of inclusion efforts, which are relatively narrowly focused; and the limited actions taken on the basis of surfaced considerations surfaced. The remainder of my dissertation focuses primarily on the dimensions of reflexivity and anticipation, for reasons discussed further in the next chapter.

CHAPTER 4: A PRESCRIPTION: DISCIPLINED, REFLEXIVE AI FUTURES

Introduction

Before delving into detail on my concrete efforts to explore possible AI futures, I briefly motivate my preliminary focus on just two dimensions of RRI: anticipation and reflexivity. I argue that the two are closely related in the context of AI futures, and that AI is relatively underdeveloped in these respects relative to other domains of technological governance. I contrast AI with energy governance, noting the much clearer articulation of assumptions, scenarios, uncertainties in the latter context. These discussions inform my use of expert elicitation, scenario planning, and formal modeling, which I briefly summarize here and then discuss and apply further in the following three chapters.

A Proliferation of AI Futures

A key characteristic of contemporary discourse about AI is disagreement about rate of current progress, future progress, and societal implications. Views about these issues abound. Telling examples of this are frequent efforts to marshal the expertise and prestige of AI researchers to discount, or vindicate, certain claims, such that human-level or superintelligent AI poses problems worth worrying about today, or that such concerns are a distraction, and writers regularly claim to speak with authority for the discipline (Etzioni, 2016; Dafoe and Russell, 2016). Vox writer Sean Illing reached out to several AI experts for comment and subsequently wrote (2018):

> There was no consensus. Disagreement about the appropriate level of concern, and even the nature of the problem, is broad. Some experts consider AI an urgent danger; many more believe the fears are either exaggerated or misplaced.

From an insider's view of these developments, it may sometimes appear clear who is right or wrong, but to an outsider, at least, AI appears to have the characteristics of turbulence, uncertainty, novelty, and ambiguity - the "TUNA" characteristics Wilkinson and Ramirez (2016) use as diagnostic criteria for issues meriting the application of scenario planning, discussed further below. Additionally, AI experts seem to be unaware of the extent of disagreement on certain topics, believing that others agree with them more than the actual distribution of views suggests (Grace et al., 2016).

Additionally, science fiction has been home to visions of AI futures for decades (or longer if one considers technologies with AI-like characteristics before the term was coined). Wall-E, Ex Machina, the Terminator series, Star Wars, Battlestar Galactica, and Star Trek are just a few examples of wildly diverse anticipations of AI. Miller and Bennett (2008) argue for science fiction's value in fostering creative reflection on alternative technological futures, a call that has been taken up by some in the AI community who advise or collaborate with authors and filmmakers on science fiction plots.

Anticipation and Reflexivity

This proliferation of possible AI futures is relatively undisciplined, in that there is little agreement even about what the right foci are (e.g., which inputs to AI systems are most important to track [Brundage, 2016a]), what the range of plausible scenarios looks like, or how to structure such analysis. For example, some forecast the future of AI based on hardware developments (Kurzweil, 2005), some refer to convergence or divergence in expert opinion (Grace et al., 2017), some explain recent AI progress by reference to the accumulation of data or specific algorithmic developments (Martinez-Plumed et al., 2018), and some explicitly bound their future knowledge at a certain time in the future, evoking a “fog” model with exponentially increasing uncertainty after a few years (Hinton, 2014). I contrast this diversity of views below with the more disciplined state of futures thinking in the energy case.

As the Grace et al. (2017) AI expert survey suggests, there is also a gap in expert awareness of the range of opinions in their field. The survey data shows views ranging from near certainty about highly negative impacts to certainty about highly positive impacts of AI (Grace et al., 2016). Anticipation of possible futures and reflexivity are closely related. While researchers/engineers (Fisher et al., 2006) and policy-makers (Brundage and Bryson, 2016) continually make decisions that influence AI outcomes, decisions will (it is hoped) be more effective if they are made with awareness of relevant uncertainties (Morgan, 2017) and what the impacts of one’s decisions might be. While AI futures cannot be predicted in detail accurately, there is likely some room for improvement by clarification of assumptions, awareness of disagreement, and updating based on evidence as events conform (or not) to one’s explicit assumptions. That is, uncertainty is not eliminable in the case of AI futures, but it can be managed and more explicitly reflected on by those involved in AI innovation. The case of energy is illustrative here.

The Comparison to Energy

The debate around renewable energy, nuclear energy, fossil-based energy, and climate change is often highly polarized and complex, and arguably insufficient action has been taken to address risks such as extreme anthropogenic climate change (Parson, 2017). As a result, energy could be construed as a case study that suggests human inability to solve such complex, multifaceted problems. However, a more optimistic appraisal of the energy system suggests a different conclusion, namely the value of decades of cumulative investments in data gathering, price analysis, conceptual framework development, regulatory model experimentation, etc. around the world. In short, what differentiates energy from AI is that in energy, there is explicit and regular analysis of trends, models, assumptions, and scenarios in a way that enables structured and fruitful discussion of policies and interventions in a common framework, and this body of knowledge and practice helps navigate some of the trade-offs communities face.

For example, take the negative externalities associated with energy usage. Progress has been made on some such challenges, including addressing damage to the ozone layer, though the bulk of the challenge of confronting climate change remains. There is a rich literature on climate change scenarios (Parson et al., 2007), including their composition and how they can be most useful for policy-making, and a detailed understanding of the energy innovation system (Gallagher et al., 2012) and the history of government interventions in this area (Sarewitz, ed., 2014). The Energy Information Administration (EIA) in the United States performs detailed data gathering and analysis of price and volume trends in energy markets, and surveys of climate change experts are commonly conducted. Finally, the Intergovernmental Panel on Climate Change (IPCC) has a rigorous process for eliciting expert views, synthesizing them, and presenting them transparently despite substantial residual uncertainties. Again, this reflexivity and disciplined anticipation has not been a panacea for action on climate change and other goals of energy policy, suggestive perhaps of the relevance here of the RRI critiques above, such as on the divergence of different parties' interests. But energy policy continues to be conducted and--to a greater extent than AI policy--routinely achieves some of its well-specified goals, such as reliable electricity delivery and improved energy efficiency in cars. By and large, developed countries continue to provide relatively cheap and widely available electricity and oil, managing uncertainties around future supply/demand through policy and investment decisions. One can find sensible analyses that take into account uncertainty, present it fairly, and iterate on prior work (e.g., US Department of Energy, 2017). Anticipation in energy is thus fairly disciplined, in part due to the long timeframe over which energy has been a critical priority for governments.

In contrast, AI futures analyses are often created de novo, with little reference to prior work, and there is no reference class of scenarios against which analysts can compare their claims. In short, there is nothing remotely resembling an IPCC or EIA for AI, an organization or process that surfaces, explains, and compares diverging views of possible futures and systematically pursues information that would allow the updating of models and their parameters in order to better (and more reflexively) anticipate possible risks and opportunities. The case of energy suggests that, while improving the analytical capacity of the AI community will not be sufficient to ensure that AI is well governed (just as the climate remains in crisis), more disciplined anticipation might be a necessary condition of effective AI governance. Analogously, it would be even more difficult to imagine progress on climate change today if anticipations of climate and energy futures were as unwieldy as in the AI case. Consider the case of AI development timelines, for example. Grace et al. (2017) find that experts disagree substantially about when certain technical capabilities will be achieved. While there are also disagreements and definitional issues in the context of energy and climate change, there is more than speculation to rest on--there are also systematic expert elicitation processes, quantitative models, and decades of historical data collection and analysis.

Anticipatory Failures

Some examples of anticipatory failures in AI illustrate the need for more attention to this dimension of RRI. China’s rise in AI (Ding, 2018), for example, received very little attention in the literature on AI governance before the past few years, and yet now threatens to reshape the entire conversation.

The rise of deep learning was largely unanticipated outside of the “Canadian mafia,” a small group of researchers largely funded by the Canadian Institute for Advanced Research (CIFAR) to push their work forward in the early 2000s and working outside the limelight in the decades prior. To illustrate the depth of this failure, consider the contributions to “AI--The Next 25 Years” (Stone and Hirsh, eds., 2005), a collection of writings marking the 25th anniversary of AAAI. The only reference to neural networks in the contributions was a passing reference to them as one among many machine learning tools. And yet, today neural networks are so pervasive in industry and cuttingedge academic research as to sometimes be mistakenly conflated with the entire field of AI by journalists.

As discussed later, some specific malicious uses of AI such as “deep fakes” (the automatic generation of artificial videos featuring a chosen person’s face) were not anticipated in advance by the developers or discussed widely in the relevant academic community in the run-up to the relevant papers’ publication, despite what would seem to be a real possibility of doing so if more effort had been devoted to anticipation.[27] Additionally, note that the core claim here is not that futures were not predicted in precise detail--this is neither possible nor a prerequisite for governance. Rather, salient properties of possible futures should at least be anticipated as plausible, and hedged against, prevented, brought about deliberately, etc. and in many cases this has not occurred when it appears that more, or more creative, effort might have done so. Subsequent developments in the field of AI, namely debate in 2019 around OpenAI’s staged release[28] of their GPT-2 system.

[27] The adult entertainment industry is often an early adopter of emerging technologies, and this is true of AI in particular, and also on the front lines of combating their misuse - see e.g. Cole, 2018a.

[28] As defined and discussed in Brundage et al. 2019 and Solaiman et al. 2019, staged release involves the release of increasingly powerful versions of an AI system over time, as opposed to releasing the most powerful version all at once. Staged release was developed for the case of GPT-2 in order to maintain option value (avoid irreversibly releasing the most powerful versions prematurely) as well as to gain information (regarding the risks and benefits of early versions of the system). Staged release is applicable to and potentially helpful in some contexts of AI where the development of some system or system component is predictably costly, e.g. in terms of computing power. It is less obviously relevant to e.g. instances of algorithmic innovation. Since the GPT-2 system was primarily distinguishable from earlier approaches by scale rather than a novel algorithm, some degree of coordination was possible among actors with more computing power than others. See Solaiman et al., 2019.

Expert Elicitation, Scenario Planning, and Formal Modeling

Surveys of AI experts, while largely restricted to expectations of the future (Grace et al., 2017) rather than prescriptions for what to do (cf. Chapter 6 below), are one step in the direction of distilling disagreements about AI in an effort to govern the technology better. But there is a long way to go before AI meets the higher, though imperfect, standards of methodological rigor and self-awareness of energy, or even other comparable technologies such as biotechnology which also have more advanced anticipation apparatuses. Below, I briefly summarize three ways in which AI governance might be more responsible by virtue of more disciplined and reflexive anticipation. The following chapters describe initial explorations and lessons learned from each. My emphasis on methods for anticipation and reflection follow from the GPT-oriented framing of AI, and help pave the way for more responsible decision-making in AI governance.

Expert elicitation, in the context of science and technology governance, refers to the systematic extraction of comparable views from one or more experts on the state of the art and possible futures of a given field, and can be done with more or less interactivity, formality, scale, etc. depending on the purpose of the elicitation (Morgan, 2014; Morgan, 2017). Surveys represent a very limited form of expert elicitation, in that they are not interactive and may not elicit comparable views when there are deeply diverging underlying models and beliefs that inform predictions. My exploration of expert elicitation went beyond most prior work by engaging experts in an interactive discussion on a focused topic (malicious use of AI), elicited feedback on detailed, explicit, near-term futures as opposed to ill-defined future states such as “human-level” AI, and was governance-relevant in that it was used to inform normative recommendations for action. I discuss this further in the following chapter.

Next, scenario planning, commonly used in the public policy and business contexts (though cf. technology governance-related applications such as Keeler, 2014), is the organized production of plausible or probable (Ramirez and Selin, 2014) futures in order to inform the evaluation of potential decisions (Wilkinson and Ramirez, 2016). Scenario planning is most useful when when grounded in a specific decision-making context and extended in space and time such that participants have the opportunity to reflect on their assumptions, change their views in response to competing perspectives, and analyze the coherence of a set of possible or plausible scenarios. Unlike prior work that focused primarily on technical uncertainties (Mankiya et al., 2017) in AI futures, or which did not comment in any detail on their methodology for the production of future scenarios (Stone et al., 2016), in chapter 5, I describe the execution of a scenario planning exercise focused on AI. This event used a variant of the Oxford Scenarios method (Wilkinson and Ramirez, 2016), and later I share lessons learned from this exercise.

Finally, formal modeling has long been used in science and technology policy analysis (Morgan, 2017; Lempert et al., 1999) in light of its ability to elucidate, more so than is otherwise possible, the specific beliefs underlying future scenarios, and to iterate and test the robustness of policy interventions across a range of scenarios. Formal modeling is complementary to approaches such as scenario planning that largely are expressed in qualitative, narrative terms (though often backed by formal modeling which informed them), and is particularly useful in the case of AI: it can explicitly account for actors with divergent interests acting rationally through the lens of game theory, and explicate the unstated assumptions underlying qualitative analyses (e.g., Bostrom, 2017) on issues with empirical dimensions such as openness in AI. In Chapter 6, I explore the utility of formal modeling for addressing some of the limitations of the above approaches, and find that it is useful in surfacing areas for further research that can address blind spots in more qualitative analyses.

Finally, note that my choice of anticipatory methods below, which largely target experts, are informed in part by the concentrated nature of AI. There is only a fairly small community of researchers (in the thousands) and policy analysts (in the hundreds) focused on AI, with a wider range of engineers and casual policy observers or practitioners with wider portfolios. In light of this unequal distribution, and the poor state of AI knowledge in broader public (Royal Society, 2017), I concentrate my initial efforts on understanding and informing expert views on the future of AI, though such efforts are not intended to suggest that public engagement is not a priority. Indeed, as I have discussed elsewhere (Brundage, 2016b) and in the prior chapter, this is valuable and could be done more rigorously, but is beyond the scope of my present contributions.

Conclusion

AI has seen the proliferation of many visions of the future. This diversity has played out in science fiction, with, e.g., Wall-E, Battlestar Galactica, and Ex Machina as just a few recent examples. A range of futures have also been suggested in public debates between tech leaders and academics. While much of the disagreement and uncertainty regarding AI futures may be irresolvable, there is likely room for better understanding of the relevant disagreements (reflexivity), and perhaps even better decision-making if assumptions are better mapped, critiqued, and tested. The following chapters build on the RRI framework and this discussion of limitations in existing AI governance discourse by attempting to more systematically explore possible AI futures.

CHAPTER 5: SCENARIO PLANNING

Introduction

Originally pioneered by Royal Dutch Shell for corporate decision-making purposes (van der Heijden, 2005; Selin, 2007), scenario planning is a method for thinking rigorously about varied futures. Its applicability beyond the corporate context has been increasingly recognized, and it is considered by many to be part of the "toolbox" of responsible innovation (Stilgoe et al., 2013) as well as science and technology policy analysis (Morgan, 2017). But its applicability to AI in particular has been less appreciated. In this chapter, I first describe characteristics of AI futures that make scenario planning a promising approach for encouraging thoughtful AI governance analysis; then I review some elements of the scenario planning literature that bear on RRI; next, I describe the scenario planning workshop I organized in June 2017, including the process of scenario construction and the substantive outputs; and finally, I discuss lessons learned both from the June 2017 workshop related to scenario planning's applicability to AI and its complementarity with other methods.

The Contestedness of AI Futures

As previously discussed above, practitioners and outside observers imagine a wide range of possible futures for AI and its social context. Even on relatively constrained questions, such as the current and future rate of technical progress, or the automatability of specific jobs, opinions vary widely. On more abstract or wide-ranging questions such as the overall societal impact of AI, opinions also diverge significantly, with views ranging from certainty that AI’s impact will be overwhelmingly positive to near-certainty that AI’s impact will be overwhelmingly negative (Grace et al., 2016).

This deep disagreement is prima facie suggestive of the value of scenario planning for AI governance: scenario planning is an approach to analysis for decision-making that takes seriously the multiplicity of plausible futures, and can inform robust decision-making given that uncertainty (Lempert et al., 1999). But the case for scenario planning for AI governance analysis goes deeper. First, scenario planning can also foster a common vocabulary for discussing alternative futures, which, as previously noted, is present and adds some value in some other domains such as energy and climate change. In particular, the unstated assumptions that different actors make about the future can be clarified and critiqued in a scenario planning context, which is particularly critical given deep uncertainty and disagreement about AI’s plausible and desirable futures. Second, scenario planning (in at least some of its forms) emphasizes the distillation of facts and perspectives into narratives, which can be more accessible to readers and listeners than abstract claims (Gong et al., 2017)--indeed, some have argued that the rise of narratives played a key role in human evolution (Boyd, 2017) and is a core component of human nature (Gottschall, 2013; Bruner, 1986). Finally, AI has other characteristics that have been argued to justify the application of scenario planning, namely turbulence (AI is progressing rapidly), uncertainty (discussed above), novelty (or at least perceived novelty--cf. Brundage and Bryson, 2014), and ambiguity (e.g., regarding what counts as AI). These “TUNA” characteristics have been used to refer to other domains such as energy and climate change (Wilkinson and Ramirez, 2016), but arguably AI is even more

"TUNA"-like in some respects, given the less developed state of analytical tools for anticipating its future--see "The Comparison to Energy" above.

It is perhaps surprising, then, that there has been little explicit effort to apply scenario planning to AI. There are a few partial exceptions which I review here before discussing the methodology, and the workshop I organized, in more detail. A common theme in prior literature on AI scenarios is that, while their substantive outputs sometimes bear resemblance to the outputs of "traditional" scenario planning exercises, the processes involved often differ, or at least aren't well described. Commonly, researchers have merely described the results of their analysis, but it is unclear how they came up with the scenario dimensions used. Another element of scenario planning practice, emphasized by many researchers (van der Heijden, 2005; Wilkinson and Ramirez, 2016), is the need for scenarios to be developed for a specific decision-making purpose and context. In contrast, one often sees more "free floating" AI scenarios in the literature. As emphasized in the scenario planning literature, the process of scenario construction is often more important than the scenarios produced.

Of the scenario-related discussions in the literature, three seem most relevant. First, the McKinsey Global Institute (Mankiya et al., 2017) has carried out several analyses of the impact of automation on the workforce and on economic growth rates in the future. In recognition of the uncertainty surrounding such issues, the authors explicitly accounted for two dimensions of uncertainty--technological progress and the rate of adoption, both of which in turn are affected by various underlying factors. Choosing a lower and higher estimate for each of these two dimensions resulted in a 2x2 matrix of possible futures, a common approach in scenario exercises (Wilkinson and

Ramirez, 2016). Second, authors at the Pardee Center for International Futures (Scott et al., 2017) also explore AI futures quantitatively, though with their dimensions of variation focused more on technical parameters and less on adoption. Their scenarios, the Current Path, Accelerated AI, and Stalled AI depict different ways in which AI could progress in the 21st century. In each of these first two cases, the authors note that they consulted relevant experts, but do not seem to carry out the sort of interactive scenario building process called for in the scenario planning literature. Third, there is a cluster of publications on uncertainties related to the future of AI, which, similarly to the above, are light on process-oriented details. Examples in this vein include Bryson's (2015) exploration of the different implications of AI depending on policy interventions related to privacy and the support of cultural diversity, and whether moral patiency[29] is attributed to AIs (Bryson, 2018); Walsh's discussion (2017) of different ways in which AI progress might stall short of "superintelligence" (e.g., the "fast thinking dog" argument); Bostrom (2014) and related writings in the same vein on different rates of AI development, making, e.g., arguments for "slow," "medium," and "fast" takeoff of AI capabilities; and Brynjolfsson and McAfee's (2014) exploration of different rates of economic changes viz-a-viz technical AI developments.

Scenario Planning: A Tool for Anticipation and Reflexivity Under Uncertainty

In the context of responsible AI governance or responsible research and innovation, scenario planning can be seen as a tool for more disciplined anticipation of

[29] In philosophy, moral agency relates to an entity's status as a maker of moral decisions, and moral patiency refers to whether the entity is entitled to moral consideration by others.

possible futures, and more regular and open-ended reflection on one’s role in the innovation ecosystem. In this section, I elaborate on the connection between the theory and practice of scenario planning, on the one hand, and the RRI dimensions of anticipation and reflexivity, on the other.

Scenario planning in something like its modern form is often traced to the work of Pierre Wack, who has been described as a "founding father" of the practice (Selin, 2005). Wack was responding to the turbulent environment that Royal Dutch Shell faced in the 1960s and 1970s. Advocates of scenario planning often emphasize the practical value that Shell's use of the technique delivered--namely, a lack of paralysis in the face of the 70s oil crisis and its aftermath, since such risks and opportunities were considered explicitly in advance (van der Heijden, 2005). Subsequently, those working in the Wack-influenced tradition have extended the theory and practice of scenario planning in myriad ways. It has been implemented in a variety of public and private contexts (Wilkinson and Ramirez, 2016), been subjected to a variety of empirical tests (see, e.g., Gong et al., 2017), and has seen conceptual development (e.g., Ramirez and Selin's [2014] discussion of scenario planning's treatment of plausibility and probability, and Ramirez and Selsky's connection of scenario planning to social ecological theories).

From an RRI perspective, the goal of scenario planning should not be to achieve parochial goals (as the technique was originally developed for), but to improve outcomes or processes of inclusion for large swathes of society. Scenario planning can be used to increase the robustness of governance analyses and decisions by scrutinizing them against a wider range of assumptions. And scenario planning can provide a vehicle for the incorporation of new stakeholders in innovation systems, as well as providing understandable outputs that can be "consumed" by those interested in imagining possible futures more clearly. In each of these respects, scenario planning can be seen as a way to improve the quality of anticipation (both by those involved directly in decision-making and by society more broadly) and as a means of increasing reflexivity (by broadening the range of outcomes, risks, and uncertainties considered).

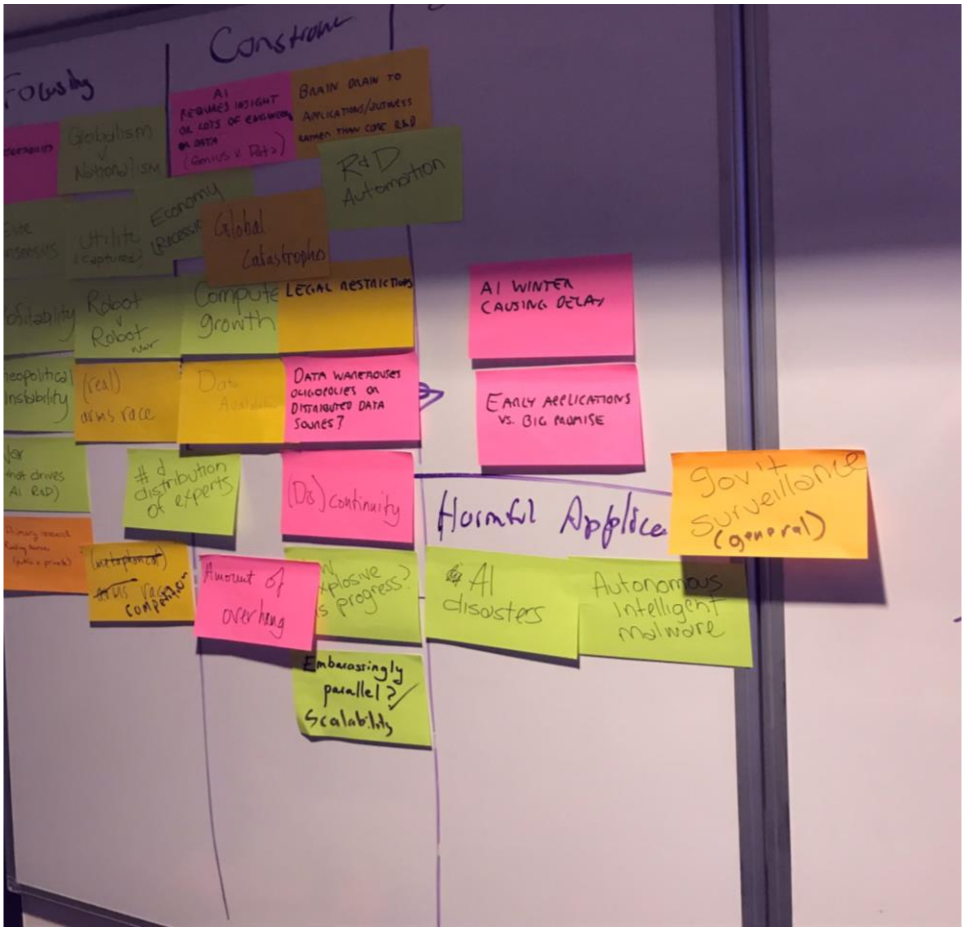
June 2017 Scenario Planning Workshop

In June 2017, I ran a workshop in Oxford, England, with 11 participants (not including myself), during which we discussed various uncertainties related to the future of AI and constructed a 2x2 matrix based on dimensions identified as particularly important and worthy of discussion. This workshop was informed by prior input from my advisors and especially Lauren Keeler (who co-facilitated the discussion), and surveys were administered to participants before and after. The workshop also built on some prior work by one workshop participant who prepared a short list of possible scenario dimensions for consideration. In this section, I elaborate on the participants, the process of running the workshop, the driving hypotheses, its outputs, and some of the feedback I received.

The 11 participants spanned a range of educational backgrounds (from some college to completed PhD) and ages (from under 20 to over 60). It was non-diverse in a number of respects, including gender (the group was predominantly male), race (the group was predominantly white), and expertise (it was predominantly an expert rather than lay person group). While some effort was made to increase diversity, I largely selected people on the basis of their involvement in discussions around AI and its future, as well as geographic proximity. The intervention was thus very limited with respect to the RRI dimension of inclusion.

The workshop was inspired by a fairly standard way of thinking about scenario planning (Wilkinson and Ramirez, 2016), in that it aimed for a 2x2 matrix of scenarios, featured extensive discussion prior to selecting the scenario dimensions, and was highly interactive. With respect to the oft-cited goal of scenario planning of facilitating an ongoing learning process, the workshop was limited in that there was minimal follow-up (just a post-workshop survey and one-off discussions with specific participants, discussed further below). An additional limitation is that there was much more time devoted to discussing scenario dimensions than fleshing out actual scenarios, leaving the final outputs fairly skeletal. As previously noted, scenario planning is best done in the service of a specific goal, and in this case, the stated goal of the exercise was to inform future research by those involved and affiliated institutions, including the Future of Humanity Institute, where the event was held.

The workshop proceeded as follows: first, those who had not filled out the pre-workshop survey were given time to do so. Then Keeler and I presented on the motivation for and process of scenario planning in general and this workshop in particular, emphasizing some of the points discussed earlier in this dissertation regarding disagreements and uncertainties. Then a workshop participant presented a slide featuring many possible dimensions for discussion, including technical, societal, and more abstract factors related to the future of AI. Extensive and fairly open-ended discussion ensued, which Keeler and I guided, with the aim of identifying dimensions of possible futures that were worth discussing further and building scenarios around. Criteria for possible scenario dimensions included their uncertainty, importance, and the ability to be influenced. Notes on whiteboards, Post-It notes, and graphs of possible scenario dimensions were used to stimulate thinking. An image from this phase of the workshop is shown below. Post-It notes like those depicted were rearranged in various configurations as the discussion moved toward selecting two dimensions around which to develop scenarios.



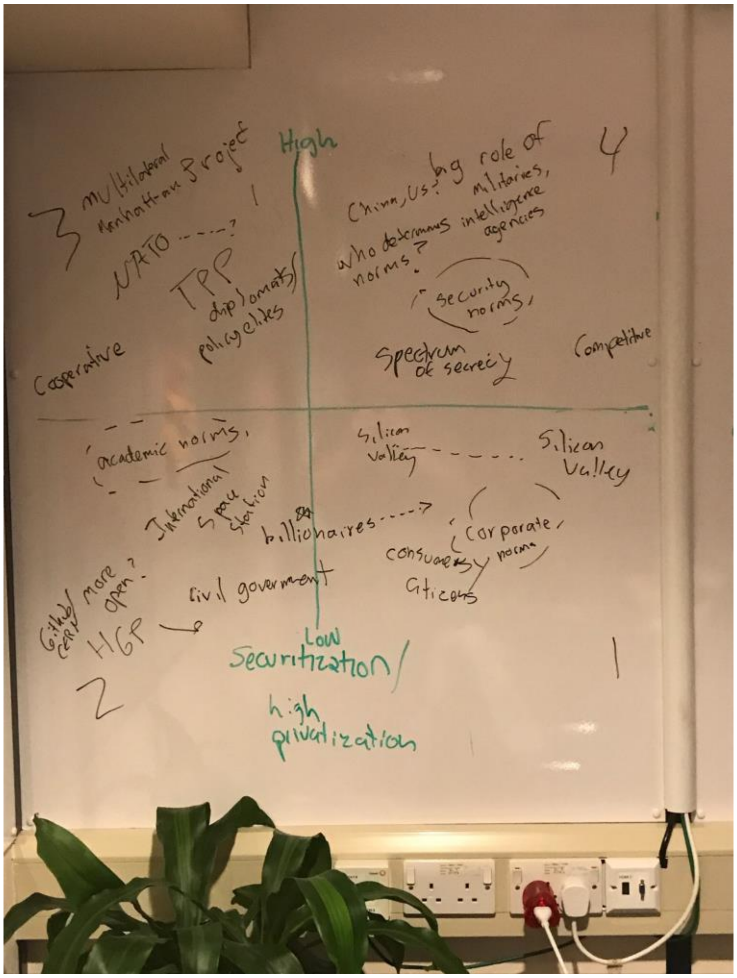
Scenario dimension brainstorming phase of workshop.

Ultimately, we selected our two dimensions for subsequent discussion. Along one dimension, we distinguished competitive from cooperative scenarios for AI development.

A more competitive scenario might involve, e.g., more closed research, government policies oriented toward attracting and retaining AI talent in one's jurisdiction, and the pursuit of applications for parochial purposes (e.g., domestic surveillance or company profit maximization). Cooperative scenarios might involve, e.g., a "CERN for AI" (discussed by various AI researchers in the past year such as Slusallek [2018]) or another style of cooperative enterprise between countries, more collaborations across national borders, and applications focused on creating global public goods.

Along the second dimension, we distinguished between securitized and unsecuritized development of AI. By securitization, we referred to the extent to which AI was seen as, and treated as, security relevant by relevant stakeholders, especially governments. In a more securitized world, a government might seek to classify much of AI research, develop AI primarily for military purposes, and set up strict export control agreements. In a less securitized world, there may be fewer constraints on publication of research, malicious use of AI may be deliberately contained or for some reason never seriously pursued, and governments are largely focused on non-security applications of AI such as for health and economic growth. Sometimes securitization was explicitly juxtaposed with "privatization" (another concept from earlier discussions), with high securitization and low privatization going together and vice versa. Notably, the competitive/cooperative and securitized/unsecuritized dimensions have sometimes been implicitly conflated in prior discussions of AI governance (as evinced in our discussions when we teased out the distinction between the two), making their explicit differentiation and exploration in a 2x2 fashion particularly fruitful.

After scenarios were selected, participants populated the different quadrants of this 2x2 matrix with various stylized facts, analogies (e.g., the "CERN for AI" concept), and risks/opportunities. An image capturing much of this discussion is shown below. This brainstorming process was used to seed the breakout groups that followed. As can be seen in the image, besides CERN for AI, a range of non-AI references/analogies were made as we discussed the range of possible governance models for AI. The image below shows notional placement of the Human Genome Project (HGP), the Manhattan Project, and the International Space Station into this 2x2 matrix.



Notes on the 2x2 scenario matrix from workshop.

Groups of 3-4 people each subsequently fleshed out scenarios built around each quadrant of this matrix. Groups were encouraged to name their scenarios, to further populate their worlds with analogies and stylized facts, and to write up their results. Some excerpts from these writings are shown below. I extracted relevant excerpts from the Google Docs created by each group, and arranged them so as to be as comparable as possible across scenarios. Only the words in italics were added by me.

High Securitization, High Cooperation *Name*: “Collective Security”/”AI Arms Control” “Drivers … for cooperation: malware/ransomware… big medical gains possible… nasty attack by small actors… [social] Disruption as threat” “More differentiation in whether to open research in a given area” *Significant actors include* “Big tech players and big governments. International organizations. (Less academic universities.) … People involved in military-complex... People in international world. Diplomats and policy elites.”	*High Securitization, Low Cooperation* *Name: no name given* *Possible actors include* “China … US … Western allies (Does the West stay together?) … Five Eyes? [Western intelligence agencies] … NATO” “Types of competition [include] cyber-conflict … drone warfare” “Nationalisation within countries” “Export controls, travel controls?” “Less rogue actors! Any basement project viewed as a security threat. Strong surveillance”
Low Securitization, High Cooperation *Name*: “Moderate AI Progress for the Public Good”/”Nice Happy-Clappy World” *Analogies* to “CERN, Human Genome Project, Academia” “Companies competitive in terms of developing AI-enabled products, but co-operative in terms of research”	*Low Securitization, Low Cooperation* *Name:* “The Market for AI” “*x* as a service will [continue to] be a slogan for everything” “If data can be collected about something, it will be automated” “VCs / investors will continue to be an important actor type”

“Agents are taxed; fair wealth distribution … Due to fair wealth distribution, global and social conflicts on the decline”	“Cooperative strategies (open source, github) continue to be important *competitive* forms of cooperation” [*original emphasis*]

Figure 3: Excerpts from scenario sketches from scenario planning workshop.

Finally, these scenarios were briefly presented to the group. Unfortunately, due to the long amount of time it took to hash out the scenario dimensions, insufficient time was taken to convert the stylized facts, analogies, and observations from discussions into actual narratives, or to have cross-scenario discussions. This point was made by some participants in response to the event and is an obvious area for improvement in future iterations.

Lessons Learned

Several observations can be made about the process and results of this workshop. First, some participants noted that the workshop was useful in thinking differently about their work (marking an at least partial success with respect to reflexivity). One participant wrote that they “had a better appreciation for China after the workshop.” Two participants noted a broadening of their foci, framed in two different ways: one “started to consider a wider variety of scenarios” and another’s views on the future of AI “became considerably more decentralized in interest.” However, not all found it equally useful, with one participant writing that “I definitely learned things relevant to the future of AI, but I'm not sure how it might have impacted my net preferences regarding the future of AI.”

Prior to the workshop, I recorded several hypotheses about the utility of scenario planning for AI governance analysis. I reproduce the hypotheses relevant to our current discussion below.[30]

"H1: Participation in the workshop will ... show less confidence in expected futures, relative to responses before the workshop. ...

H3: Participation in the workshop will result in more concern about misuse scenarios, relative to responses before the workshop.

H4: Participation in the workshop will result in more new policy-related ideas identified than the model condition."

As elaborated in footnote 16, not all of these could be confirmed or rejected with the data obtained, but I will make a few brief comments on them here.

Some of the comments reported above bear on H1--in further discussion with some participants, it seems that the workshop was useful in encouraging participants to think concretely about the role that the Chinese government in particular, and governments more generally, would and/or should play in the future of AI.

[30] Not all of the hypotheses are relevant because of statistical power issues related to the small sample size (11 participants and one intervention), and the fact that the "modeling condition" discussed below did not ultimately involve the same survey questions. But I include them all below for completeness and transparency. Note that H1 was modified above due to a typo in the original version below. They were all sent to Lauren Keeler in advance of the workshop.

"H1: Participation in the workshop will weaken result in survey responses that show less confidence in expected futures, relative to responses before the workshop.

H2: Participation in the workshop will result in survey responses that show less confidence in expected futures, relative to participants in modeling condition.

H3: Participation in the workshop will result in more concern about misuse scenarios, relative to responses before the workshop.

H4: Participation in the workshop will result in more new policy-related ideas identified than the model condition.

H5: Participation in the modeling condition will result in more confidence in expected futures, relative to responses prior to the modeling intervention.

H6: Participation in the modeling condition will result in more new empirical research ideas, relative to the workshop condition."

Quantitatively, however, there was no sign of an effect on confidence. The mean level of confidence before the workshop was 58.75, and afterwards it was 58.59, where 0 represents choices made effectively at random and 100 represents full confidence. This is not a substantial difference, even if the sample size were large enough for statistical significance, which it wasn't.

With regard to H3, discussion of misuse scenarios, for participants who had not thought in any detail about such issues before, may have encouraged reflexivity. Unfortunately, the data was insufficient to show any significant effect of the workshop on confidence levels. And it is unclear whether the workshop resulted in more policy-related ideas or research questions (H4) than alternative possible interventions, but it certainly resulted in some such ideas that individual participants hadn't previously considered. Some examples of future research areas found in the notes produced by participants include the distribution of computing hardware production (relevant to competitive and/or securitized scenarios and their associated geopolitics), the prospects for "trusted third parties" to mediate conflicts between states, and the utility of zero-knowledge proofs (Kroll et al., 2016) to enabling accountability for AI systems without necessarily revealing all details of the systems when there are legitimate privacy or security concerns.

More broadly, the flow of discussion and the workshop's output was consistent with the general RRI principles discussed earlier--for example, much attention was paid to the research and development process but also to broader (and often unintended) social implications of AI, which is appropriate for a general purpose technology. Additionally, some misuse risks were discussed which are not widely discussed in the literature, or at least not previously known to the participants brainstorming them (e.g., the use of AI to make hyper-addictive digital assistants or games), indicating the wide variety of possible societal outcomes of AI.

Another test of the workshop's utility is whether policy analyses developed elsewhere, without consideration of these scenario dimensions, are robust to the risks discussed at the workshop. If they are not, then this is a sign of scenario planning achieving one of its intended purposes: surfacing lack of robustness in possible plans. To test this, I consider two influential analyses of AI governance: the White House's "Preparing for the Future of Artificial Intelligence" report (Executive Office of the President, 2016) and Calo's "Artificial Intelligence Policy: A Primer and Roadmap" paper (2017). I specifically examine whether a key feature of the scenario dimensions, the possible securitization of AI over concerns about malicious or military uses, is considered. In both publications, the word "terrorist" and its variants do not appear, and "crime" or "criminal" and variants only appear in two specific contexts: criminal justice applications of AI, and war crimes related to autonomous weapons. However, the cybersecurity implications of AI are briefly considered in each case. "International cooperation" and related terms were also searched for in each document. There is some discussion in the White House document of the value of engaging with other countries on AI, but little that would shed light on specific ideas considered at the scenario planning workshop such as a global "CERN for AI" or multilateral efforts to govern access to computing hardware.[31] These are just two examples of relevant publications which could

[31] Note that a CERN for AI has been discussed previously (e.g., ITU, 2017; Slusallek, 2018) but not explored in much detail or compared against competing multilateral governance architectures like those discussed in the scenario planning workshop, to my knowledge.

be considered, but based on my familiarity with a wider range of literature, I do not believe that examining other publications would substantially change the picture: the state of understanding on AI governance is widely considered to be incomplete (as each of the above documents agrees) and one should not expect to find very substantial policy prescriptions for managing all possible risks and opportunities. Given the short duration of the scenario planning workshop, and the surfacing of issues that are not well explored in the literature, it seems that we are far from the point of steeply diminishing returns on this tool for analysis.

The scenario planning workshop was perceived by at least some participants as useful, as was the process and output of the "Bad Actors and AI" workshop, but they were useful in different ways. Whereas the expert elicitation discussed earlier around malicious use was much more focused, the scenario planning workshop was more about identifying crucial considerations and developing a vocabulary for discussing possible futures. Additionally, whereas the expert elicitation around misuse focused in large part on specific instances of misuse, the scenario planning workshop focused more on the broader context and relative importance of different actors. The participants noted that in some scenarios, actors like intelligence agencies would be more important, whereas in others, corporations or academics would be. This points to a fundamental challenge of AI governance which had not (to my knowledge) been explicitly addressed previously in the RRI literature, and was not explicitly a part of any of my writing on AI governance prior to the workshop. Namely, actions taken by certain actors today may influence the extent of both their, and others', future power over events. The securitization of AI may result, for example, in a decline in academics' influence relative to governments. This is a key consideration for RRI, and while it may not be entirely novel, it was one of my main takeaways from this experience and will inform some of my future research.

Conclusion

Scenario planning is a decades-old practice oriented toward grappling with uncertain futures and reflecting on implicit assumptions. As such, it behooves AI governance researchers and practitioners to seriously consider its utility in grappling with their own uncertain futures of interest. The June 2017 workshop highlighted some of the logistical and representational challenges associated with performing such exercises, e.g., allowing appropriate time for fleshing out scenarios and ensuring a diversity of participants. It also identified some substantively interesting anticipatory outputs that aren’t well addressed in existing analyses, and was perceived as useful in some respects by participants, apparently increasing their reflexivity. Future work could extend this pilot in various ways, such as broader discussion of the downstream societal implications of the different scenarios, as well as greater efforts to surface the root causes of participants’ different perspectives on these topics. An additional method that complements these approaches is discussed in the next chapter: expert elicitation.

CHAPTER 6: EXPERT ELICITATION

Introduction

Expert elicitation is a method for extracting and synthesizing assessments of the likelihood of future scenarios from the tacit knowledge of experts. The technique has been used in a range of domains, including both private and public sector decision-making (Morgan, 2014; Morgan, 2017) and has seen only limited application to AI to date. In this chapter, I describe two cases in which I have applied expert elicitation, with varying degrees of formality, and discuss lessons learned from these cases.

Expert elicitation is a powerful tool for systematically extract information about technical trends, societal impacts, and governance affordances. As noted in Chapter 2, not all possible applications of AI can be anticipated, but some can, and understanding underlying trends in capabilities can provide a valuable input into broader anticipatory work. Additionally, understanding the range of normative opinions, as I do in the second case studied here, can provide a key knowledge base for informing governance decisions.

Expert Elicitation

Slottje et al. (2008) define expert elicitation as “a systematic approach to synthesize subjective judgments of experts on a subject where there is uncertainty due to insufficient data, when such data is unattainable because of physical constraints or lack of resources.” Expert elicitation is not applicable for all problems, at least in the formal version sometimes employed, where experts’ subjective views are converted into a form that is comparable across individuals, such as probability density functions, or PDFs

(Solttje et al., 2008) representing expectations about future events. As Morgan (2014; 2017) notes, a key question to ask before using expert elicitation is whether there are relevant experts. This is a key question in the context of AI, and informs my use of expert elicitation discussed below.

The only case to date of probabilistic estimates being elicited by experts (to my knowledge) is in a recent series of surveys about the future of AI (Müller and Bostrom, 2016; Grace et al., 2016; Grace et al., 2017). A question on these surveys would ask, for example, the year by which one believes there is a 10% probability that AI will exhibit human-level[32] performance on a particular task, or on all economically relevant tasks. Or, instead, the question may be framed in the opposite way: given some year, what probability does one assign to that level of performance. Interestingly, respondents give very different answers based on these otherwise equivalent framings (Grace et al., 2016). Additionally, several other interesting trends have been observed, such as that North Americans and Asians have highly different expectations of the future, with Asians expecting AI to develop more quickly (Grace et al., 2017). And both within and across different regions, there is substantial heterogeneity, with some expecting rapid development toward human-level performance across all economically relevant tasks in the coming two decades, and others expecting a more prolonged development process lasting centuries (Grace et al., 2017).

[32] Note that "human-level" performance is not an unproblematic benchmark, even in the context of a given task, given diversity across humans. Additionally, AI systems often learn and reason in different ways from humans (e.g., learning less efficiently in simulation, versus humans who learn more efficiently in the real world, leveraging language).

Other less formal instances of expert elicitation include consulting of experts on analyses of the future of labor automation (Mankiya et al., 2017) and more general reports on AI over different time frames, such as the “AI in the year 2030” report by the Stanford 100 Year AI Study (Stone et al., 2016) though typically the precise ways in which experts were consulted and aggregated is left vague.

For the purposes of the explorations below, it is worth noting a few limitations of expert elicitation suggested by the governance-related properties of AI. First, as discussed in Chapter 3, the generality of AI makes accurately forecasting all specific uses of the technology intractable, though some successes are possible in cases where a clear demand for an application can be foreseen. Second, the state of explicit modeling of AI progress is limited (Brundage, 2016a), making it difficult to convert answers to specific survey questions (e.g., on the rate of progress in hardware or algorithms) into a coherent model of different AI development scenarios. Third, expert elicitation in its most intensive forms (Morgan, 2014; Morgan, 2017) involves substantial feedback and iteration, such that experts are made aware of a wide variety of relevant considerations, have the opportunity to reflect on their confidence levels, and so forth. But as noted in the literature (Morgan, 2017), it is not always worth this investment of time and resources, as only certain issues are amenable to getting useful information from (formal) expert elicitation. As discussed below, I extend expert elicitation in AI into the area of eliciting views on normative questions, and into new empirical areas (misuse of AI), but do not reach the frontier of methodological rigor as shown in other technological domains--see, e.g., Usher et al., 2013 for a more formal and interactive methodology used for answering a specific question related to climate change. Fourth, expert elicitation is primarily aimed at surfacing rather than creating new tacit knowledge: not all governance-relevant questions about which one might want information have been thought about much by experts, and expert elicitation excels at extracting information that is already somewhat clear in an expert's head but has simply not been converted into explicit knowledge in the technical literature. In contrast, the scenario planning method used in the previous chapter focuses more on creating new knowledge about dimensions of possible futures which had seen little prior explicit analysis.

Finally, and especially importantly, expert elicitation in this case has been applied primarily to AI experts and adjacent AI-interested experts about specifically AI-related matters. Insofar as AI is applied to a wide range of industries and sectors of society, a huge range of expertise ought to be brought to bear in informing responsible governance for AI. The cases studied here, then, barely scratch the surface.

Bad Actors and AI Workshop

In February 2017, Shahar Avin of the Centre for the Study of Existential Risk and I co-chaired a workshop entitled "Bad Actors and Artificial Intelligence," which involved three dozen participants representing a range of expertises such as cybersecurity, machine learning, counterterrorism, AI safety, and autonomous weapons. The purpose of the workshop was to surface the expertise of these participants, combine it into a coherent account of the landscape of risks related to deliberate misuse[33] of AI, as opposed to unintended harms such as bias or safety accidents. Here I describe the genesis and

[33] In the ultimate report, the term "malicious use" was used rather than misuse, since the latter would seem to include the unintended risks discussed above (e.g., bias). However, here I use the two terms interchangeably as sometimes the word "misuse" alone is more succinct.

process of that workshop, the report-writing process that followed, and how a wide range of centrally involved and external experts' input was brought to bear on the conclusions of the report.

Before the workshop, in 2016, experts were informally surveyed about their interest in a prospective workshop in a number of different areas. This feedback ultimately informed the scope of the final workshop, "Bad Actors and AI." Alternative possible workshop topics included the intersection of cybersecurity and AI and international AI cooperation. The reason the ultimate focus ended up being "Bad Actors and AI" was that a range of experts concurred on the importance of this topic, found cybersecurity to be relevant to AI misuse but not exhaustive of the range of risks, and there was generally agreed to be little useful literature on the topic to date. As discussed in the final report (Brundage and Avin et al., 2018), there have been many discrete discussions of specific forms of AI misuse, and calls for attention to the topic, but as yet little in the way of a comprehensive analysis or discussion of possible preventative or mitigation measures.

Several initial experts were then invited to participate in the workshop, and a snowball process was used to recruit people with useful expertise in cases where the initial invitees were unavailable. In addition, feedback was solicited both from workshop participants and non-workshop participants on a framing document authored by myself and Avin, which laid out a range of prospective misuse types. This feedback informed the structure of the workshop and provided early refinement of our sense of which risks were plausible or implausible.

The workshop itself, held under Chatham House rules, consisted of two days of content. On the first day, experts presented a range of past- and future-looking views on issues ranging from terrorist weaponization of drones to automation of cybercrime. Breakout groups then explored such issues in more detail, with an eye toward increasing focus on the most important risks, which were captured in the form of handwritten notes. From an expert elicitation, both the presentations and the discussions were rich opportunities for extracting tacit knowledge, as most speakers were discussing topics that had never previously been discussed in this particular way. For example, one speaker discussed the various contributors to AI progress and development, including talent, hardware, data, and software, which had previously been discussed to some extent in the literature (Brundage, 2016a), but in this case, the speaker focused on analyzing such elements from the perspective of misuse prevention, i.e., where the points of control are for a society concerned about misuse. The breakout groups, too, elicited participants' views on which risks from a vast landscape of potential concerns were actually worth worrying about.

On the second day of the workshop, focus shifted toward prevention and mitigation measures. An agreement was reached that more could and should be done about misuse, but that writing a research agenda would be a useful output from the group. At this point, the most formal form of expert elicitation used for the workshop took place: participants were surveyed about which interventions against misuse (out of those surfaced in the discussion) they considered to be tractable and useful. Information from this surveying process was, in turn, used to prime the writing process. At a meta-level, this information was also useful in identifying areas of controversy: for example, there was near-unanimity on some interventions, and deep controversy over others. This elicitation aimed to surface both normative ("useful") and empirical ("tractable") opinions from the group.

While writing the report, the team of authors circulated drafts of the report to dozens of experts in related fields. For example, experts on formal verification provided useful input on the tractability of confirming certain properties of AI system that may reduce misuse risks (e.g., susceptibility to adversarial examples). We learned a lot about what sorts of recommendations would be controversial in the AI community: namely, those related to openness, which informed how we presented such topics, though ultimately the lead authors accepted some risk of controversy in light of the stakes involved. In the terms of the RRI framework, these reactions raise further concerns (in addition to those discussed in Chapter 2) about responsiveness in the AI community: while speculation about misuse seems to be within the AI community's Overton window ("the range of ideas tolerable in public discourse"--Wikipedia, 2018), anything that would change academic norms such as those related to publication are seen more skeptically. This may in fact be justified, and the report did not call for any radical changes (but rather, reflection and expectation), but highlights the need for further attention to the recalcitrance of norms and whether they are rightly or wrongly calcified.[34]

After releasing the report on February 20, 2018, one year after the workshop, there was substantial attention paid to the report in the media and AI community, which provided further insight into how these topics are seen by experts. To date, the main

[34] Cf. arguments elsewhere for more openness and reproducibility in AI, e.g., Islam et al., 2017.

report website has received over 70,000 visits, and this probably significantly understates the number of total downloads since the report is available in other places, too. So there will likely be a long period in which the corresponding authors (myself included) will receive inquiries about the subject of the report, and already this has shed some light on the sorts of risks people find worrying. For example, there has been substantially more attention paid to the cybersecurity and "fake news" elements of the report, perhaps pointing to the forward-looking orientation of the cybersecurity section, on the one hand, and the pervasive existing concern about fake news, on the other. Synthetic media in particular has subsequently become a focus of substantial policymaker scrutiny in a range of countries.

When OpenAI announced its language modeling system, GPT-2, in February 2019, the initial blog post (which I co-authored) noted the potential for the system to be misused by generating large-scale disinformation. Subsequent research by OpenAI and partnering organizations extended this analysis, finding that GPT-2 outputs were capable of deceiving people, and that with some engineering effort, malicious actors could steer the system toward generating propaganda (Solaiman et al., 2019). In Solaiman et al.'s report (in which I was closely involved as second author), we found that statistically distinguishing machine-generated from human-generated text would be a difficult challenge, and pointed toward a number of steps that could prepare. The short timeframe over which the connection between AI and fake news became decision relevant for AI developers suggests the value of conducting regular expert elicitation.

Scenario Planning Workshop

The scenario planning workshop I conducted in June 2017 had elements of expert elicitation in addition to exploring scenario planning methodology; I discuss the expert elicitation dimensions of that workshop here.

In June 2017, I organized a workshop with Lauren Keeler devoted to exploring possible AI futures. The methodology of the workshop was informed by the Oxford Scenario Planning Approach (Wilkinson and Ramirez, 2016), and the scenario planning aspects of this event are discussed further in the next chapter. Here, I focus more on surveys conducted before and after the workshop, which shed light on the views of this small (n=9)[35] group of AI experts on the empirical and normative aspects of AI.

The surveys administered before and after the workshop were identical except that the after ("post-test") version included some questions about the utility of the workshop. 37 questions were asked, ranging from basic demographic information to variations on previous questions about AI futures (Grace et al., 2017) to novel questions about openness in AI. Appendix A lists the questions asked in the pre and post surveys. Several results stand out.

First, some of the results were broadly consistent with larger n studies of expert opinion in AI. The median expectation of time until broadly human-level AI was also mid-century (with a mean of 2048), similar to Grace et al.'s findings (2017), although in this case the participants were drawn from a different distribution. While Grace et al. used participants who had published at the NIPS and ICML conference series, the

[35] 9 respondents responded to both the pre- and post-test surveys.

researchers participating in the scenario planning workshop mostly had not (owing to a greater policy focus in their research output or being at an earlier stage in their careers). This suggests a fairly pervasive view of “a few decades from now” as a focal point for expecting more advanced AI, in the core AI research community and adjacent communities, though here too there was substantial variation in views.

Second, while a statistical power analysis conducted before the workshop suggested that we were unlikely to see statistically significant effects of participation in the workshop, the results were nevertheless interesting and novel in some respects. Similarly to how views on empirical matters are widely distributed, it was also true in this sample that experts varied widely on their views on normative questions.

For example, experts differed greatly on their preferred level of openness in the AI community in the future: when asked how open the AI community should be in the future, 2 respondents said “completely open,” 1 said “mostly open,” 2 said “an even mixture of open and closed,” and 4 said “mostly closed.” In response to another survey question, 5 respondents characterized current AI capabilities as “highly concentrated,” whereas 3 said “an even mix of distributed and concentrated,” and 1 said “mostly distributed.” Respondents largely agreed on empirical questions related to prevailing norms in the AI community, with all but one respondents viewing the AI community as mostly open today (the remaining respondent viewed the AI community as fully open). There was more diversity of views regarding how these features of the AI ecosystem should or could change in the future. When asked how concentrated or distributed AI capabilities should be in the future after further technical progress, 4 said “mostly distributed," 2 said "an even mix of distributed vs. concentrated," and 3 said "mostly concentrated."

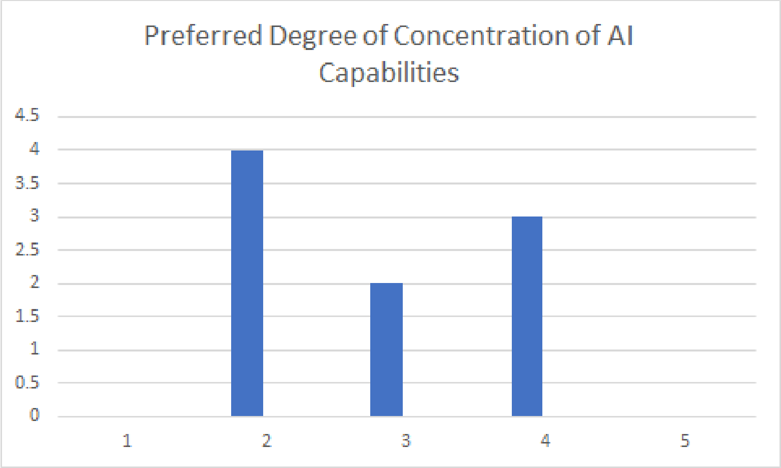


Table 1: Preferences regarding the degree of concentration in AI capabilities from pre-workshop survey. The Y axis corresponds to the number of people who chose the option. The X axis is arranged from "Fully distributed" (1) to "Fully concentrated" (5).

The above findings are highly preliminary but also are, to the best of my knowledge, the first to suggest the following plausible hypothesis: AI experts generally agree on the state of norms in the community today, though differ over their preferred future for it as well as their expectations for how it will evolve in the future. That finding was true both about openness issues as well as the role of governments in future AI development.

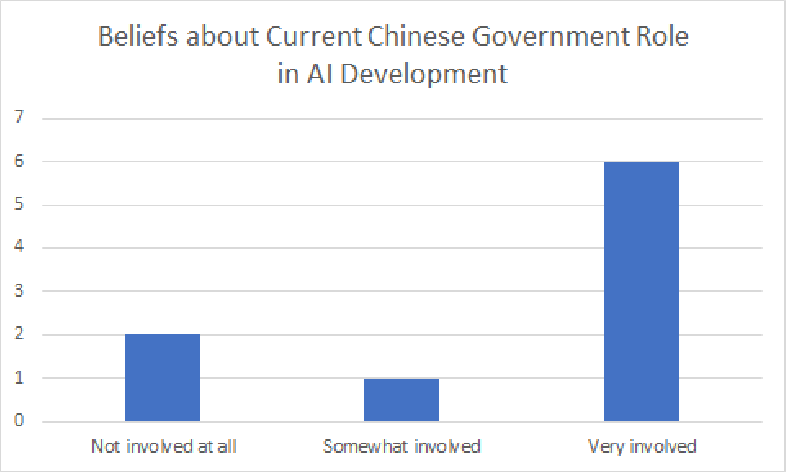


Table 2: Views on the current involvement of the Chinese government in AI development, from pre-workshop survey. The Y axis corresponds to the number of people who chose the option and the X axis reflects the 3 options given. See Appendix A for full questions.

Similarly, out of the 9 participants who answered the question "How involved do you think the United States government is today in AI research, development, and deployment?", 7 answered "somewhat involved" and 2 answered "very involved." None answered "not involved at all." Respondents also expected greater involvement in the future (with the number of "very involved" increasing to 5, and "somewhat involved" reducing to 4), but differed again over which outcome they preferred.

Lessons Learned

Several lessons can be derived from the interventions discussed above.

First, similar to other technical domains, there does seem to be substantial tacit knowledge of a variety of governance-related aspects of AI that aren't surfaced in the written literature. This includes information related to misuse potential and about the prevailing norms in the community. Eliciting views on these topics was successful in that surprisingly few major disagreements were found about descriptive features of the world (e.g., whether certain misuse scenarios were plausible in 5 years), suggesting the value of future expert elicitation exercises over a wider range of domains and with different motivating questions. However, further investigation will be required in order to assess whether this agreement reflects surfaced, accurate tacit knowledge or overconfidence. It is also possible that AI or its malicious use will not proceed as quickly as anticipated by many experts, if (for example) a general “many things are possible in a few years” mindset has pervaded the community in light of recent successes. While I am generally confident in the findings of the report (Brundage and Avin et al., 2018), and feel that caveats were used appropriately (e.g., about the possibility that applications will be technically possible but not actually realized due to different foci and motivations of actors or preventative measures), the report was an exploratory exercise and as such carries some risk of being misguided in various respects.

Second, interdisciplinary research proved to be critical in the “Bad Actors and AI” workshop. The risks of malicious uses of AI involve issues that cannot be resolved solely by experts in the AI community, and it is fortunate that such a diverse array of experts were willing to participate in the workshop. A critical question for future exploration is: which areas of expertise have been insufficiently tapped in order to shed light on AI futures and AI governance? For example, in later work related to the publication of GPT-2, experts in online abuse, disinformation, bias, and a number of other areas contributed to the decision-making process (Solaiman et al., 2019).

Third, the difficulty of anticipation in AI, given the technology's generality, can be seen in the aftermath of the misuse report. While the general category of fictitious media generated by AI was anticipated in the early version of the report, it was not until near the time of publication that the deep fakes phenomenon began to surface on the Internet. The term "deep fakes" originates from a Reddit user who began posting videos of celebrities' faces swapped onto pornographic videos, and other users began to apply the technique to a wide range of celebrities and non-celebrities. Earlier writers had anticipated something broadly like deep fakes (Weninger, 2015) but at the workshop, we did not consider some scenarios that subsequently came to pass, such as the use of deep fakes for "revenge porn" (Cole, 2018b). The report did, however, highlight the risks of natural language generation, a concern that now appears more salient in light of GPT-2 and other recent language models.

Conclusion

Expert elicitation has substantial potential for informing responsible governance of AI, via improving anticipation and reflexivity, and in particular it excels in providing a structured way to surface knowledge from a range of experts. In the two cases discussed here, there was value added by the technique in surfacing preexisting tacit knowledge on governance-related issues. Expert elicitation shed light on normative disagreements among experts as well as remaining substantial disagreements on the future of AI, especially over the long-term. The next chapter explores the complementary value of expert elicitation as a tool for anticipation and reflexivity.

CHAPTER 7: FORMAL MODELING

Introduction

Formally modeling the dynamics of a technology and its societal implications is a long-standing aim of many researchers, with advantages and disadvantages (Morgan, 2017). In this chapter, I briefly discuss the merits of formal modeling as a tool for governance analysis, and then describe two collaborative efforts I have undertaken toward this end. First, I describe a model of openness in AI inspired by the agent-based modeling paradigm. Second, I describe preliminary efforts to analyze competition over AI development using the tools of game theory, which have subsequently been extended in published work. I conclude with some lessons learned from both of these projects.

Formal Modeling

As suggested in Chapter 4, formal modeling has various attractive properties from the perspective of the RRI framework. Computers can represent and manipulate models with greater precision and speed than humans, and as such they can explore the implications of a given set of assumptions to their logical conclusion. In this respect, computers are synergistic with human capabilities (Lempert et al., 1999): humans are better at intuiting causal models, suggesting plausible ranges for variables, etc., and computers can carry these through to their conclusion.

Given this setup, models can then output conclusions either for a single set of parameters, or across a wide range of possible parameters. While not a panacea, “running” models across a wide range of possible parameters can partially alleviate concerns about the accuracy of such models. When the goal is not to perfectly capture a system but to explore a range of possible system dynamics, results can be somewhat more trustworthy (though care must still be used in interpreting model outputs). Similarly, here I do not attempt to make crisp predictions about the future of AI but to capture salient properties of possible system dynamics.

Several frameworks and literatures have utility for exploring technology-related futures with formal models. Systems dynamics models (Morgan, 2017), agent-based models (Blume, 2015), and game theoretic models (Maschler et al., 2013) have all been used for governance analysis. Bespoke models for climate change (Parson et al., 2007) have also been developed, combining economic, ecological, and other factors. Similar efforts have taken place in other policy domains. Here, I focus in particular on agent-based models and game theoretic models, described further in the sections below.

Modeling AI Openness in an Agent-Based Modeling Framework

Agent-based modeling, or ABM, focuses on creating “computer programs in which intelligent agents interact in a set environment based on a defined set of rules” (Theodorou et al., 2019). Key subjects of investigation using the agent-based modeling paradigm include the evolution of cooperation and the diffusion of information through social networks.

In my exploration of the utility of formal modeling for analyzing AI governance, I considered a range of possible governance-related issues of concern, including malicious uses of AI and international cooperation. I first sought to explore the potential of agent-based modeling to help understand the dynamics of openness in AI. Openness is a much-discussed topic in AI and has come up in contexts such as reproducibility (Islam et al., 2017) as well as various issues related to governance (Calo, 2011b; Krakovna, 2016; Bostrom, 2017). Further, in 2019, openness in AI development became a major topic of debate in the aftermath of OpenAI's partial release of GPT-2. Given the close connection between issues of cooperation, conflict, and communication in the agent-based modeling paradigm on the one hand, and discussion of these issues in the discussion of openness in AI, on the other, using ABM to study openness in AI appears to be a good fit.

To explore the possible dynamics of openness in AI in a quantitative way, I collaborated with David Kristoffersson[36] to develop an agent-based model of openness in AI using the Python package Mesa (Masad and Kazil, 2015). Mesa provides a set of tools for representing and automating simulations of multi-agent dynamics, and is well-maintained and documented relative to other analogous code bases we considered leveraging. In particular, we explored the connections between the initial distribution of AI developers (i.e., their skills and resources), the manner by which AI developers do or don't share their research with one another, and the final distribution of AI "capabilities" after some period of further research and dissemination. The modeling we did in Mesa[37] did not ever result in any crisp conclusions or recommendations, but the process of formalizing intuitions about openness was illustrative of the many uncertainties surrounding the topic, and shed light on areas for possible future research. Here, I describe the range of assumptions which were inputted to the model, and some overall

[36] Kristoffersson contributed to the software engineering side of this project, and also played a key role as a thinking partner on the specification of system properties to be engineered. He has not been involved at all in the writing of this chapter, and does not necessarily endorse the conclusions herein.

[37] An iPython notebook with associated code can be found here: https://github.com/dkristoffersson/aidevsim/blob/master/aidevsim.ipynb

takeaways from the process. Throughout the development process for this model, early versions were discussed with other researchers in order to iterate the underlying assumptions, but the project has still not achieved a sufficient degree of maturity for widespread sharing, in part due to the difficulties encountered.

Following the affordances of the Mesa package, the model[38] involves a set of agents with various properties, including their initial capabilities and their preferences. Capabilities were at first represented as a single number. After further analysis and reflecting on some previous qualitative work on openness (e.g., Bostrom, 2017), capabilities were disaggregated further into fixed and shareable capabilities. Fixed capabilities might include, e.g., capital to invest in hardware or skill of developers. Shareable capabilities might be particular algorithmic insights or codebases. The latter are more easily shareable than the former, though in practice the line between fixed and shareable capabilities is blurred. Various possible initial distributions of capabilities were representable in the model (e.g., a normal distribution or a power law distribution of each type of capability, and parameters for tweaking the details of these distributions).

Next, different types of sharing and technological progress were considered. Sharing might be represented in various ways, and as was discovered during this project, little had been said about this in written discussions of the issue, so new formalisms had to be invented.

[38] In the context of Mesa, a model could refer to a particular assignment of parameters (e.g., number of agents, dynamics of interaction, etc.) which is simulated. Here, I use model to refer to the overall architecture of the program, within which many “models” in the former sense could be simulated.

For example, if two AI developers, A and B, share all of their capabilities openly, and their capabilities at that point in time could be quantified as 10 and 20 respectively, what is the resulting level of capabilities for A and B? 30 is one naive answer, but this seems unlikely as talent and hardware also seem to matter, such that the uptake of capabilities by all actors is not 100%. Additionally, there is substantial overlap in what different developers know (due both to drawing on a common literature as well as simultaneous invention), and research may pertain to various distinct dimensions of AI progress as opposed to a single dimension. These and other considerations led to a wide range of "sharing types" being modeled, such as "Leader" (rising to the level of the highest capability shared), "Proportional gain from all" (each gains some from capabilities shared by others, but to different extents), and "All" (each rises to the level of the sum of shared capabilities). Each set of assumptions led to different final distributions after the AI developers shared their research. Additionally, during the course of the model stepping forward in time (for some pre-set number of steps), new research would likely occur. This, in turn, was represented through various possible growth functions (linear, exponential, sigmoid, etc.).

Finally, given these initial endowments and subsequent sharing (or lack thereof, if some agents are designated as sharing with only certain other agents or with no one), some final distribution is calculated. But it is not clear which, if any, governance-related metric ought to be applied to these final distributions. I used the Gini coefficient (a measure of inequality) as one way of quantifying the effects of openness dynamics on societal outcomes, but a Gini coefficient is not the only possible means of such analysis. In visualizations of the data, the maxima and minima of capabilities were also considered, as well as averages. Additionally, since the preferences of the agents were represented as a single continuous dimension, which could stand in for, e.g., maliciousness, safety-awareness, beneficence, etc., this information could be visualized alongside the resulting capability distribution. So one could see, for example, the results of agents preferentially sharing their research with others who share similar values.

Notwithstanding this effort to enumerate, formalize, and simulate the results of a range of possible assumptions, the model was never polished enough to widely distribute or comprehensively analyze. The failure of the project to move toward completion is partly a result of the complexity and uncertainty surfaced in the process of formalization: a wide range of parameters were incorporated, to the point that the model lacked elegance or clear plausibility. As suggested above, robustness could in principle be increased by running the model for a wide range of parameters. While I did this to an extent, a common failure mode was that many of the results were prima facie implausible. Capabilities would often skyrocket (or not) due to the interplay of various factors, and glitches in the computation of capabilities were not uncommon in the development process (e.g., when capabilities that were shared communally were also unintentionally added back to one's own capabilities).

Despite these practical and conceptual problems, the agent-based modeling project did provide some value through its failure. It made clear the limited state of explicit understanding of openness dynamics in AI and increased my own reflexivity about openness in AI. To my knowledge, several of the design choices that were made were resolutions of "problems" that had never even been posed previously (e.g., the many possible sharing dynamics discussed above). Insofar as openness is a key parameter of future AI governance (Brundage and Avin et al., 2018; Solaiman et al., 2019), there is a need for a renewed exploration of these uncertainties, though it is as yet unclear whether the agent-based modeling paradigm is the right one. Additionally, note that some of the issues raised by my openness modeling--how, if at all, to quantify the capabilities of different actors, and understand the relative usability of shared information to different actors--have become more pressing in the context of GPT-2's publication, and arose explicitly in the associated analysis (Solaiman et al., 2019).[39] Real-life instances of complicated openness decisions suggest the possibility of revisiting agent-based modeling in the future in light of empirical data that could be used to constrain the space of scenarios more. Below I discuss follow-up work done collaboratively with Askell and Hadfield on game-theoretic aspects of AI development, which pushed this interest in formal modeling of AI development scenarios further.

Game Theory as a Lens on an "AI Arms Race"

Game theory (Maschler et al., 2013) is a formal framework for studying the choices made by rational actors when their fates (their "payoffs" in game theory terminology) are dependent on each others' actions. Game theory studies optimal decision-making in interdependent contexts, and is thus relevant both to characterizing contemporary AI governance as well as identifying potential changes to incentives ("payoffs") that could move the world into a more favorable type of game. Unlike agent-

[39] Specifically, in the report by Solaiman et al. on "Release Strategies and the Social Impacts of Language Models" (for which I am second author), we noted that actors vary widely in their technical skill and motivation to misuse language models. Consequently, we organized our analysis along the lines of distinct sets of actors.

based modeling, game theory tends more to look for “exact” solutions and predictions for a limited range of assumptions, rather than considering a wide range of possible scenarios and observing the aggregate statistics. Game theory’s study of interdependent decision-making has been highly influential in economics, public policy, international relations, and other areas. I am not aware of any instance of its explicit application in the RRI literature,[40] but it speaks directly to some of the RRI limitations discussed previously, namely the conflicts of interests between different stakeholders in innovation. And in the broader literature relevant to RRI, instances of game-theoretic reasoning applied to science and technology governance are more widespread. For example, “races to the bottom” on regulation are sometimes discussed wherein a competitive economic environment drives behavior that is bad for all involved, an instance of the kind of collective action problem discussed here in the context of AI.

Game theory has been particularly fruitful in the area of international relations, where scholars such as Thomas Schelling pioneered the application of game theory to understanding, predicting, and managing international conflict. Ideas of deterrence and mutually assured destruction, for example, developed in the Cold War, owe a strong debt to game-theoretic reasoning (Kaplan, 1983). Here, I discuss an exploration of game theory applied to competition between countries developing AI, which in recent years has come to be referred to by some as an “AI arms race.” Citing the many military and economic applications of AI, many have claimed, for example, that the US and China are or are entering an AI arms race (see, e.g., Simonite, 2017). Others have pushed back on

[40] A search of the *Journal of Responsible Innovation*, for example, resulted in no hits for “game theory.”

this narrative, noting that an arms race is both a dangerous framing of the situation and not the only way to think about international relations related to AI (Cave and ÓhÉigeartaigh, 2018). Here I go further, and suggest an alternative framing of prospects for AI competition and cooperation, grounded more explicitly in game theory.

Armstrong et al. (2016) are the first, to my knowledge, to formally analyze a prospective AI arms race, and they come to a troubling conclusion: under certain assumptions, states or companies competing to develop AI might "defect" by skimping on safety investments, resulting in worse outcomes for all. Arguably, this is an accurate characterization of what is happening today with driverless cars, where many jurisdictions and companies are "racing" to be leaders in the area. This race to deployment has already caused several deaths, and could threaten the future of the industry as a whole by negatively coloring public and policymaker perceptions of the technology.

Armstrong and colleagues implicitly frame the problem of AI competition as a Prisoner's Dilemma, a well understood archetypal "game" (decision-making context) studied by game theorists. The term Prisoner's Dilemma is inspired by a canonical example of two prisoners being interrogated and tempted with incentives to "rat out" their co-conspirators in the adjacent room. Since each knows that the other will be similarly tempted, they both end up betraying the other and end up worse than if they could have credibly committed to cooperating with one another by staying silent. Prisoner's Dilemma-inspired problem framings yield pessimistic predictions: in the absence of means to coordinate and with incentives to defect, both sides will make decisions that result in lower overall utility than if they could have coordinated in some fashion.

Not all archetypal games yield such pessimistic predictions, however. The outcome of a game depends on the actors, their interests, and the actions available to them. In the context of AI governance, RRI requires attention to existing incentives, as well as an openness to changing those incentives in order to move into a more favorable game.

Here, I discuss an alternative framing of the AI cooperation/conflict question and give a more mixed prediction, namely that it is possible for states to avoid a dangerous AI arms race (where e.g., autonomous weapons and automated hacking systems escalate in speed and scale to a dangerous level) if they can find ways to signal their intentions. Instead of a Prisoner's Dilemma, I frame AI competition/cooperation as a Stag Hunt Game. In a Stag Hunt, two or more parties choose between hunting for easy prey (a hare) or hunting for more difficult prey (a stag). Each would prefer that both hunt for the stag, which yields more meat (the "payoff dominant equilibrium"), but each is also uncertain about what the other will do, and can safely ensure a meal by hunting the hare (the "risk dominant equilibrium").

The payoff matrix below depicts this general framework in the context of a symmetric game. Cooperation is predicted to occur if the payoffs are ordered $a > b \geq d > c$, where the first letter in each row corresponds to the payoff for the row party, and the second letter corresponds to the payoff for the column party.

	Cooperate	Defect
Cooperate	a, a	c, b
Defect	b, c	d, d

Figure 4: Generic 2x2 game matrix.

A Stag Hunt framing doesn't yield a crisp prediction of what choice the players will make. Multiple equilibria are possible, but under the right circumstances, both parties will gain from cooperation. A Stag Hunt is thus a much more desirable game to "play" than a Prisoner's Dilemma.

Is AI development a Stag Hunt or a Prisoner's Dilemma?[41] The answer could have significant implications for the prospects for responsible AI development.

Simplifying a lot, this question reduces to whether it's better for both sides if they both cooperate, and this in turn can be broken down further into variables such as the direct and indirect risks of cutting corners or attacking one another, the size of the bounty attainable through AI, and each side's views on the long-term downsides of the other side attaining an advantage.[42] Below, I'll briefly suggest some reasons why the Stag Hunt framing deserves serious consideration in the context of AI governance analysis. Anecdotally, analyses like this have been well received to date, with one such researcher writing: "Hadn't considered asking whether it's PD [Prisoner's Dilemma] or SH [Stag Hunt], and asking this seems to move my mind toward thinking about different

[41] It could be neither, of course--this is only a partial and suggestive analysis. A wide range of games have been explored in the game theory literature, and there has been little application of game theory to AI development to date.

[42] These points are formalized in Askell et al., 2019, discussed below.

questions." Such a statement could be seen as an improvement in the reader's reflexivity and anticipatory repertoire, a clear win from an RRI perspective.[43]

Several lines of reasoning suggest that states might want to cooperate rather than defect in some cases. Regarding the potential upsides of cooperation, consider first the scalability of AI (Brundage, 2018a) in virtue of its status as a digital technology. There is no reason in principle why AI cannot be simultaneously possessed by many parties, and provide economic and other gains to many countries simultaneously. Additionally, there are possible means by which AI could be developed jointly, if efforts were made in advance to arrange this (e.g., a global "CERN for AI"). Additionally, there is an open question as to how bad it would be for one country to be behind another in AI development. This in turn depends in part on whether AI is an offense-dominant or defense-dominant technology (Glaser and Kaufmann, 1998). For example, it could be that even if another state has more advanced AI than you, either in general or in in a particular domain such as cybersecurity, you remain able to defend yourself due to other advances such as geography (long distances tend to favor defenders, as projecting power over a long distance is difficult even with contemporary ICTs). Finally, there is some repetition involved in this game (i.e. not all decisions will be made instantly tomorrow, but there are opportunities for reactions to play out over time). Repetition in the game of AI developments means that states might anticipate their defection leading to a worsening of the race, and therefore not do so in the first place.[44]

[43] Note that I focus here on reactions my earliest work on the topic, though subsequent co-authored (Askell et al., 2019) was eventually published and more widely discussed.

[44] Note that a Prisoner's Dilemma with repetition is an Iterated Prisoner's Dilemma, another classic game type with greater prospects for cooperation. The Stag Hunt framing I use here implicitly collapses this "shadow of the future" into a single stage.

This preliminary analysis is far from conclusive, but it gives some reason to suspect that the Prisoner's Dilemma model is not obviously the only possible framing of international AI development. If, for example, states can take steps that signal their intentions for developing AI, they may be able to move toward cooperative equilibria.

Later work by Askell, Brundage, and Hadfield (2019) elaborated on some aspects of these issues, extending the analysis to inter-corporate competition and making additional theoretical and practical points. The authors compare AI safety to other areas such as car safety, and highlight the importance of minimal standards to prevent a regulatory "race to the bottom." The paper derives five factors that are correlated with the likelihood of cooperation in a particular instance: high trust, shared upside (from cooperation), low exposure (to defection from others), low advantage (from defecting), and shared downside (from mutual defection). Notably, these factors are all defined with respect to the beliefs of AI developers, highlighting the critical importance of narratives regarding AI "races." In my contributions to this paper, I highlighted policy steps that flow from this game-theoretic picture of AI development. I argued that four steps can increase the real and perceived alignment of different parties' interests, given those variables. These four steps are: promote accurate beliefs about the benefits of cooperation, collaborate on shared research and engineering challenges, open up more aspects of AI development to appropriate oversight and feedback, and incentivize adherence to high standards of safety. While much work remains to be done on this topic, formal modeling is beginning to enable more granular analysis of the assumptions and causal reasoning underlying different claims about AI futures.

Conclusion

Formal modeling of AI governance remains in an early stage. I have taken some early steps here, but much more remains to be done. The agent-based modeling project never reached maturity, and the Stag Hunt formulation of AI competition and cooperation, while formal and explicit about its assumptions, is not yet quantitative or fully researched. Indeed, it may be that much further progress is not yet attainable at this time in light of the lack of empirical data to “fix” certain parameters of these models (e.g., the offense/defense balance in AI, the AI development capabilities of different actors, the ability to credibly signal one’s benevolent intentions in AI development, or the dynamics of sharing research). But the early evidence suggests the possible utility of formal modeling as a complement to other approaches. In particular, agent-based modeling allows one to pursue various sets of assumptions to their ultimate conclusions, and game theory enables analogies to be drawn across analogous cases from history and contemporary international relations. Anecdotal evidence suggests some success in improving my own and others’ reflexivity, and increasing our analytical repertoire for anticipating AI futures.

CHAPTER 8: CONCLUSION AND FUTURE DIRECTIONS

Introduction

This dissertation has made several contributions to the theory and practice of AI governance. First, I developed a general account of AI's governance-related properties. I did this by building on the RRI framework and emphasizing distinctive properties of AI that suggest a need for particular attention to specific challenges and methodologies. I critiqued extant governance efforts through the lens of RRI, pointing to deficiencies with respect to the dimensions of anticipation, reflexivity, inclusion, and responsiveness. Motivated by the commitments underlying RRI, I proposed and analyzed the results from several explorations into AI futures using the methods of scenario planning, expert elicitation, and formal modeling. In this chapter, I take stock of these contributions in more detail, make some normative recommendations in light of the analysis, and suggest future research directions.

Contributions of the Analytical Framework

Much of the extant discussion of AI governance has taken place at the object level, rather than the meta level: researchers and others have analyzed AI and its societal dimensions, and proposed various actions to improve processes or outcomes. This dissertation has primarily analyzed such topics at a meta level instead: it critiques those efforts, and aims to improve the methods and concepts employed to think about AI governance. Leveraging the RRI framework has helped to identify issues with existing approaches and to suggest methods with potential utility in improving AI governance analysis.

This high level critique has surfaced several limitations of extant work. While some significant progress has been made to date in AI governance, there remains deep uncertainty and disagreement about possible futures, and little in the way of a common framework for grappling with these uncertainties. Reflexivity and responsiveness are far too low, and few high level principles developed in, e.g., open letters and statements of principles, have been grounded out in practice and widely implemented. My findings added ammunition to the idea that we have serious challenges ahead, noting significant normative disagreement related to openness in AI and the appropriate role of government that had not been previously discussed.

Yet there are some reasons for optimism. The limited rigor of extant efforts also represents an opportunity for future work, as tools and ideas from other domains can be brought to bear on AI governance. My discussion of RRI in AI has shed light on the nature of AI as a general purpose technology, thereby providing even more urgency to the task of anticipating and shaping AI's future as well as a pointer to a rich body of work to draw insights from (specifically other GPTs). Further, some early signals from my efforts are encouraging, such as an improved understanding of malicious uses of AI, a broadening of thinking about international competition and cooperation related to AI, and the broad consensus among surveyed experts about some features of the current landscape.

Toward an Integrated Portfolio of Anticipatory Methods

While some results from this inquiry have been encouraging, not all efforts were successful. Some merely logistical problems could be avoided in the future, such as running out of time in the scenario planning workshop. And some more serious problems such as the challenge of applying agent-based modeling cast doubt on my choices of methods for AI-related anticipation. Distilling these successes and failures, a few themes can be derived which can inform future choices of methods for AI governance analysis. In particular, I suggest that the tools discussed be used in a more tightly integrated manner.[45]

Scenario planning is perhaps the most generic of the tools explored. It has seen wide application across sectors of society and issue domains, and there is no reason to think it cannot yield additional insights in the case of AI governance. However, the way in which it was employed in this case left something to be desired, given the short duration of the exercise and the relatively limited pool of participants. The latter, regarding the limited pool, is both a problem with this particular instance as well as a general problem with the method, since it benefits from interactivity, yet its interactivity does not scale well to the wide range of people potentially affected by AI (at least absent efforts to automate the process). In the case of my scenario planning workshop, I had done substantial preparation for the event and had experienced and helped carry out scenario planning myself. I also had a much more seasoned practitioner and theorist to

[45] In some cases, I originally planned to integrate the three methods discussed here more closely in the first place, e.g., informing scenario planning with results from the agent-based model, but deficiencies of the latter impacted efforts at integration.

assist me, and such expertise is a limited resource. It is possible to scalably distribute the verbal or other outputs of such workshops, but the interactivity of scenario planning is critical to its success in revising assumptions and increasing understanding, and this process itself is not as easily scaled. Reading scenarios produced, on its own, will not suffice, so an open question for future work is where the limited resource of skilled scenario planning practitioners can best be applied, in the domain of AI governance, to maximal effect, and/or how to scale the technique better in the future. Additionally, one of scenario planning's upsides (the engagingness of narrative) is intimately tied to one of its downsides, namely, the unlikelihood of any particular scenario coming to pass and the possibility of overconfidence arising from consuming a narrow range of scenarios. This suggests the need for conducting many diverse scenario exercises so as to broaden participants' perspectives, and the complementarity of scenario planning with other approaches that can systematically model and process possible scenarios in ways that humans cannot easily process, namely with formal modeling.

Expert elicitation was found to be useful in distilling knowledge from a range of relevant experts, including but not limited to AI researchers. The resulting product (Brundage and Avin et al., 2018) was widely distributed, and was targeted not merely at experts but at a wide audience, thus scaling quite favorably in terms of outputs. However, similarly to scenario planning, the full experience of grappling with possible AI misuse scenarios cannot be as easily conveyed, and perhaps some of the value of participating in the workshop was lost in translation. Additionally, expert elicitation was useful for identifying key agreements and disagreements across experts, including on topics that had not been previously posed to them in a formal way. This is valuable for informing future work, and offers a good complement to scenario planning, where experts from one AI-adjacent area can contribute their insights, via elicitation, to scenario planners, and vice versa. But a deep problem remains, which is that some of the uncertainties related to AI are perhaps irresolvable via expert elicitation. As Morgan (2014; 2017) emphasizes, a key question in evaluating the utility of this method is whether or not suitable experts actually exist. Determining the distribution of expertise on topics such as the societal implications of AI is non-trivial. Eliciting a wide range of expert opinions and then subjecting them, after some time delay, to reexamination, as I have done with some of my own technical forecasts (Brundage, 2018b), may serve to both help create such experts (by providing them with structured feedback and allowing them to improve their statistical calibration) and surface the existence of experts (by identifying those who already outperform others). Cultivating and surfacing the insights from such experts would not suffice to settle the issue, especially with regard to normative disagreements, but would provide a stepping stone to more useful expert elicitation exercises, and, through cross-fertilization across methods, scenario planning and formal modeling exercises, as well.

Finally, formal modeling demonstrated some promise for explicitly identifying and reasoning about uncertainties related to the future of AI. While I was met with mixed success, better understanding the scope of the challenge was key to future work, both in particular domains like openness and across AI governance more generally. The fact that I was unable to find, for example, a directly transferable model of openness from another technical domain, was itself useful information. More work will need to be done in order to grapple with the complex dynamics of AI futures. With regard to game theory, some success was had but was anecdotal in nature. The game theoretic formulation of international AI cooperation and conflict helped elucidate some key properties for future examination, such as the offense-defense balance in AI. Formal modeling alone, though, will be insufficient, as determining the right model structure and model parameters has to be informed by other forms of analysis. An entirely de novo model will be of no use, and I learned from the agent-based modeling case that it is much easier to come up with possible dimensions of variation than to determine where we are likely to find ourselves along those dimensions.

One can see, then, an emerging portfolio of methods that integrate well with one another. Scenario planning, the most generic of the methods considered, can be enriched by results from expert elicitation and formal modeling exercises. The assumptions of experts, in turn, can be tested as they read scenarios that bring their beliefs into question. Other methods not explored in this dissertation, such as quantitative technological forecasting (Brundage, 2016), could also enrich each of these, by providing fodder for scenarios, questions and priors for elicited experts, and parameter assignments for formal models. None on its own will be a panacea, but this dissertation has helped to pave the way for forging this integrated portfolio and further extending the boundary of knowledge about AI governance.

Normative Recommendations

Before turning to the final section of this chapter, describing areas for future work (in addition to the methodological integration-oriented ones above), I make two specific recommendations for the practice of governance. These are aimed at addressing what I see as especially major deficiencies in extant efforts, and are targeted at major institutional stakeholders such as governments, trade groups, and companies.

First, in light of its GPT status and the deficiencies of purely market-driven decision-making (see, e.g., Sachs, 2008), AI should be more consciously steered toward socially beneficial ends by investors such as governments. As previously noted, nothing like a “GiveWell for AI” exists--that is, a systematic analysis of prospective and/or existing AI applications with respect to their societal impacts and cost-effectiveness in addressing societal problems. While AI should not seen as a straightforward technological fix for all problems (Sarewitz and Nelson, 2008), it does have many attractive features such as the ability to reduce costs of certain goods and services, and to perform tasks at a scale not previously possible. Per Sarewitz and Nelson’s analysis, technological fixes are most plausible in cases where (among other things) a preexisting technical core exists, around which the solution can be built. This appears to be the case for many possible AI applications, which can be used as apps or components of apps in widely available smartphones. As such, a sustained and substantial effort should go into identifying opportunities for such innovations, engaging stakeholders about potential undesired impacts (Jasanoff, 2016), and enacting the most promising ones, even (and perhaps especially) when they are likely to be unprofitable and therefore not pursued by default.

Second, AI governance needs to be much more effectively democratized. A GPT has pervasive applications and implications, and only a tiny fraction of affected parties have been seriously involved in discussions of governance to date. Efforts such as the White House workshops discussed earlier, DeepMind’s efforts to engage clinical stakeholders and patients, and the production of “explainer” videos by a range of parties are steps in the right direction, but don’t go nearly far enough. A serious risk of the agenda laid out in this dissertation is that, if it is only taken up by a small group of parties, group-think and elitism will predominate, and, like early applications of scenario planning in the oil industry, the benefits of better anticipation and reflexivity will accrue to the users rather than broader publics. The equity and inclusion dimensions of RRI should not be given short shrift in AI governance efforts such as the emerging intercorporate and intersectoral Partnership on AI, as well as the increased attention to AI by governments around the world. In this respect, efforts such as the NASA Asteroid Initiative should be seen as inspirations for the AI community (Tomblin et al., 2017), and we are beginning to see more ambitious efforts in the direction of public engagement on AI (Mahmood and Kaplan, 2019).

Future Directions

I conclude with some additional suggestions for future work in building a more robust theory and practice of RRI in AI.

First, some of my conclusions were limited in their definitiveness by the small sample sizes involved. It is difficult to conclude much from the expert elicitation exercise data, for example, given that statistically significant results were not able to be found. Descriptive statistics can be useful, as can anecdotes, but these are only the first step. A first area of future work, then, is merely to scale up these explorations through similar exercises across a range of different AI governance topics.

Second, the exercises employed could have been deepened in various respects. For example, the scenario planning workshop could have been longer, and rich scenarios could have been both finished and circulated back to the 3 other break out groups that didn't create them, enabling a richer discussion than occurred with only the skeletal aspects of the scenarios completed. Participants could have revisited these scenarios over time, as they continued in their work, and reflected on the ways in which elements of them did or did not come true. And the workshop could have been enriched by outputs from other methods, such as formal modeling, or a presentation of the statistics from the participants' own surveys. Future work, then, should ensure that similar exercises are given the time and resource commitments they deserve, so that the scenarios are not merely created and then dropped, but used as part of an ongoing and sustained reflection process.

Finally, I suggest ways to deepen the theoretical understanding of AI governance. If AI is being developed for domain X, the existing governance literature on X and current practitioners of X are a critical resources to consider. While this may sound obvious, it has not always been done in practice. Lipton (2017), for example, notes the uncritical assumptions made by some in the literature on machine learning interpretability for healthcare, which he argued need to be re-examined in light of health stakeholders' actual opinions. Additionally, lessons from historical instances of GPTs should be more comprehensively explored for lessons about AI governance challenges. Historically, the GPT literature has primarily consisted of economists and economic historians; it is time to take a broader, interdisciplinary perspective on this issue in light of the steady progression of a GPT in our midst today. And beyond GPTs in general, there are likely further lessons to be learned by critically examining the relationship between other ICTs in particular and AI, as these ICTs are AI's closest cousin in the socio-technological family tree. While the literature on existing ICTs and their governance was considered in the writing of this dissertation, more could and should be done to identify synergies the lessons of AI, the Internet, smartphones, personal computers, and more. As 2018 and 2018's headlines attest, with Mark Zuckberberg's Congressional testimony in the wake of data breaches currently receiving substantial attention as I first wrote this chapter, the inability or unwillingness of countries to address key issues related to privacy, security, and fairness on online platforms raise critical questions about our prospects for successfully governing this next wave of ICTs responsibly. A substantial increase in awareness, depth of understanding, and effort will be required.

BIBLIOGRAPHY

AAAI. "AI—The Next 25 Years." Association for the Advancement of Artificial Intelligence. Accessed April 15, 2018. http://www.ai.rutgers.edu/aaai25/.

AI4ALL. "AI4ALL - Official Website." Accessed March 9, 2018. http://ai-4-all.org.

Amodei, Dario and Danny Hernandez. "AI and Compute," OpenAI blog post, May 16, 2018. https://openai.com/blog/ai-and-compute/

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete Problems in AI Safety." ArXiv:1606.06565 [Cs], June 21, 2016. http://arxiv.org/abs/1606.06565.

Armstrong, Stuart, Nick Bostrom, and Carl Shulman. "Racing to the Precipice: A Model of Artificial Intelligence Development." AI & SOCIETY 31, no. 2 (May 1, 2016): 201–6. https://doi.org/10.1007/s00146-015-0590-y.

Arulkumaran, K., Marc Deisenroth, Miles Brundage, and Anil Bharath. "Deep Reinforcement Learning: A Brief Survey." IEEE Signal Processing Magazine 34, no. 6 (November 2017): 26–38. https://doi.org/10.1109/MSP.2017.2743240.

Askell, A., Miles Brundage, and Gillian Hadfield. "The Role of Cooperation in Responsible AI Development," arXiv preprint server, https://arxiv.org/abs/1907.04534

Balaram, Brhmie, Tony Greenham, and Jasmine Leonard. "Artificial Intelligence: Real Public Engagement," The Royal Society for the Arts (RSA), 2018. https://www.thersa.org/globalassets/pdfs/reports/rsa_artificial-intelligence---real-public-engagement.pdf

Barben, Daniel, Erik Fisher, Cynthia Selin, and David Guston. "Anticipatory Governance of Nanotechnology: Foresight, Engagement, and Integration" In Hackett, EJ, Amsterdamska, O, Lynch, M, Wajcman, J. (eds) The Handbook of Science and Technology Studies. Cambridge, MA: The MIT Press, pp. 979–1000.

Barnden, J. et al. "AISB-00 Symposium on Artificial Intelligence, Ethics, and (Quasi-)Human Rights," symposium website, 2000. Accessed April 17, 2018. http://www.cs.bham.ac.uk/~jab/AISB-00/Rights/

Baum, Seth. "Reconciliation between Factions Focused on Near-Term and Long-Term Artificial Intelligence." SSRN Scholarly Paper. Rochester, NY: Social

Science Research Network, May 29, 2017. https://papers.ssrn.com/abstract=2976444.

Bergen, Mark, and Jeremy Kahn. “Google AI Researcher Accused of Sexual Harassment - Bloomberg.” December 16, 2017. Accessed March 9, 2018. https://www.bloomberg.com/news/articles/2017-12-16/google-researcher-accused-of-sexual-harassment-roiling-ai-field.

Bimber, Bruce. The Politics of Expertise in Congress: The Rise and Fall of the Office of Technology Assessment. SUNY Press, 1996.. http://www.sunypress.edu/p-2413-the-politics-of-expertise-in-co.aspx.

Black in AI. “Black in AI” Accessed March 9, 2018. https://blackinai.github.io/.

Blok, Vincent, and Pieter Lemmens. “The Emerging Concept of Responsible Innovation. Three Reasons Why It Is Questionable and Calls for a Radical Transformation of the Concept of Innovation.” Responsible Innovation,Vol. 2, 2015. https://doi.org/10.1007/978-3-319-17308-5_2.

Blume, Lawrence. Agent-Based Models for Policy Analysis. National Academies Press (US), 2015. https://www.ncbi.nlm.nih.gov/books/NBK305903/.

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

———. “Strategic Implications of Openness in AI Development.” Global Policy 8, no. 2 (May 1, 2017): 135–48. https://doi.org/10.1111/1758-5899.12403.

Boyd, Brian. “The Evolution of Stories: From Mimesis to Language, from Fact to Fiction.” Wiley Interdisciplinary Reviews. Cognitive Science 9, no. 1 (January 2018). https://doi.org/10.1002/wcs.1444.

Bresnahan, Timothy F., and Trajtenberg, Manuel. “General Purpose Technologies.” Working Paper. National Bureau of Economic Research, August 1992. https://doi.org/10.3386/w4148.

Brundage, Miles. “Economic Possibilities for Our Children: Artificial Intelligence and the Future of Work, Education, and Leisure.” In AAAI Workshop: AI and Ethics, 2015.

———. “Modeling Progress in AI.” ArXiv:1512.05849 [Cs], AAAI Workshop on AI, Ethics, and Society, 2016. http://arxiv.org/abs/1512.05849.

———. “The White House AI Workshops and Public Engagement in Science and Technology,” 2016, blog post at milesbrundage.com. Accessed March 9, 2018. http://www.milesbrundage.com/blog-posts/the-white-house-ai-workshops-and-public-engagement-in-science-and-technology.

———. “Scaling Up Humanity: The Case for Conditional Optimism about Artificial Intelligence,” chapter in “Should we be afraid of artificial intelligence?,” report by European Parliamentary Research Service. 2018. http://www.europarl.europa.eu/RegData/etudes/IDAN/2018/614547/EPRS_IDA(2018)614547_EN.pdf

———. “Review of My 2017 Forecasts.” milesbrundage.com (personal blog), 2018. http://www.milesbrundage.com/1/post/2018/01/review-of-my-2017-forecasts.html.

Brundage, Miles, and Shahar Avin et al. “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.” arXiv preprint server, February 20, 2018. http://arxiv.org/abs/1802.07228.

Brundage, Miles, and John Danaher. “Cognitive Scarcity and Artificial Intelligence: How Assistive AI Could Alleviate Inequality.” Philosophical Disquisitions blog, May 2017. Accessed March 9, 2018. http://philosophicaldisquisitions.blogspot.co.uk/2017/05/cognitive-scarcity-and-artificial.html.

Brundage, Miles, and Joanna Bryson. “Smart Policies for Artificial Intelligence.” ArXiv:1608.08196 [Cs], August 29, 2016. http://arxiv.org/abs/1608.08196.

———. “Why Watson Is Real Artificial Intelligence.” Slate, February 14, 2014. http://www.slate.com/blogs/future_tense/2014/02/14/watson_is_real_artificial_intelligence_despite_claims_to_the_contrary.html.

Brundage, Miles, and Guston, David. “Understanding the Movement(s) for Responsible Innovation,” chapter in International Handbook on Responsible Innovation, (von Schomberg and Hankins, eds.), 2019.

Brundage, Miles, Alec Radford, Jeffrey Wu, Amanda Askell, David Lansky, Danny Hernandez, Daniela Amodei, and David Luan. 2019. “GPT-2 Interim Update,” OpenAI blog, https://openai.com/blog/better-language-models/#update

Bruner, Jerome. “Actual Minds, Possible Worlds.” Harvard University Press. 1987/ http://www.hup.harvard.edu/catalog.php?isbn=9780674003668.

Brooks, Rodney. “Artificial Intelligence Is a Tool, Not a Threat.” Rethink Robotics, November 10, 2014. http://www.rethinkrobotics.com/blog/artificial-intelligence-tool-threat/.

Brian, Jenny Dyck. “Special perspectives section: responsible research and innovation for synthetic biology,” Journal of Responsible Innovation, Volume 2, Issue 1. 2015.

Brynjolfsson, Erik and Andrew McAfee. “The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies,” W.W. Norton, 2014.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson. “Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics.” Working Paper. National Bureau of Economic Research, November 2017. https://doi.org/10.3386/w24001.

Bryson, Joanna. “The Intelligence Explosion started 10,000 years ago (+/- 2,000)” Adventures in NI blog, 2013. https://joanna-bryson.blogspot.co.uk/2013/12/the-intelligence-explosion-started.html

———. “Robots are more like novels than children,” Adventures in NI blog, 2015. https://joanna-bryson.blogspot.com/2015/03/robots-are-more-like-novels-than.html

———. “Artificial Intelligence and Pro-Social Behaviour.” In Collective Agency and Cooperation in Natural and Artificial Systems, 281–306. Philosophical Studies Series. Springer, Cham, 2015. https://doi.org/10.1007/978-3-319-15515-9_15.

———. “The Meaning of the EPSRC Principles of Robotics.” Connection Science 29, no. 2 (April 3, 2017): 130–36. https://doi.org/10.1080/09540091.2017.1313817.

———. “Patiency Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics.” Ethics and Information Technology 20, no. 1 (March 1, 2018): 15–26. https://doi.org/10.1007/s10676-018-9448-6.

———. “The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation,” The Oxford Handbook of Ethics of Artificial Intelligence, Oxford: Oxford University Press. 2019. Quote from author’s version, found at http://www.cs.bath.ac.uk/~jjb/ftp/Bryson19AIforLawofAI.pdf

Bryson, J. J., Y. Ando, and H. Lehmann. “Agent-Based Modelling as Scientific Method: A Case Study Analysing Primate Social Behaviour.” Philosophical Transactions of the Royal Society B: Biological Sciences 362, no. 1485 (September 29, 2007): 1685–99. https://doi.org/10.1098/rstb.2007.2061.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. “Semantics Derived Automatically from Language Corpora Contain Human-like Biases.” Science (New York, N.Y.) 356, no. 6334 (14 2017): 183–86. https://doi.org/10.1126/science.aal4230.

Calo, Ryan. “Peeping Hals.” Artificial Intelligence, Special Review Issue, 175, no. 5, 2011. https://doi.org/10.1016/j.artint.2010.11.025.

———. “Open Robotics.” Maryland Law Review, Vol. 70, No. 3, 2011. https://papers.ssrn.com/abstract=1706293.

———. “Robotics and the Lessons of Cyberlaw.” California Law Review, Vol. 103, No. 3, pp. 513-563, 2015. https://papers.ssrn.com/abstract=2402972.

———. “Artificial Intelligence Policy: A Primer and Roadmap.” SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, August 8, 2017. https://papers.ssrn.com/abstract=3015350.

Campolo, Adam et al. “AI Now 2017 Report,” The AI Now Institute. https://ainowinstitute.org/AI_Now_2017_Report.pdf

Cave, Stephen, and Seán S. ÓhÉigeartaigh. “An AI Race for Strategic Advantage: Rhetoric and Risks,” 2018.

Christensen, Henrik I. et al. “A Roadmap for US Robotics: From Internet to Robotics, 2016 edition,” roadmap organized by University of California San Diego et al. November 7, 2016. Accessed March 9, 2018. http://jacobsschool.ucsd.edu/contextualrobotics/docs/rm3-final-rs.pdf

CIFAR. “Pan-Canadian Artificial Intelligence Strategy Overview.” Canadian Institute for Advanced Research. Accessed March 9, 2018. https://www.cifar.ca/assets/pan-canadian-artificial-intelligence-strategy-overview/.

Cockburn, Iain M., Rebecca Henderson, and Scott Stern. “The Impact of Artificial Intelligence on Innovation.” The Economics of Artificial Intelligence: An Agenda, January 10, 2018. http://www.nber.org/chapters/c14006.

Cole, Samantha. “Fake Porn Makers Are Worried About Accidentally Making Child Porn.” Motherboard, February 27, 2018. https://motherboard.vice.com/en_us/article/evmkxa/ai-fake-porn-deepfakes-child-pornography-emma-watson-elle-fanning.

———. “Deepfakes Were Created As a Way to Own Women’s Bodies--We Can’t Forget That,” Vice, June 18, 2018. https://www.vice.com/en_us/article/nekqmd/deepfake-porn-origins-sexism-reddit-v25n2

Clark, Jack. “Artificial Intelligence Has a ‘Sea of Dudes’ Problem.” Bloomberg, June 23, 2016. https://www.bloomberg.com/news/articles/2016-06-23/artificial-intelligence-has-a-sea-of-dudes-problem.

Crawford, Kate, and Ryan Calo. “There Is a Blind Spot in AI Research.” Nature News 538, no. 7625 (October 20, 2016): 311. https://doi.org/10.1038/538311a.

Crootof, Rebecca. “Artificial Intelligence Research Needs Responsible Publication Norms,” Lawfare blog, October 24, 2019. https://www.lawfareblog.com/artificial-intelligence-research-needs-responsible-publication-norms

Danaher, John. “The Rise of the Robots and the Crisis of Moral Patiency.” AI & SOCIETY, November 18, 2017. https://doi.org/10.1007/s00146-017-0773-9.

Davies, Sarah R., and Cynthia Selin. “Energy Futures: Five Dilemmas of the Practice of Anticipatory Governance.” Environmental Communication 6, no. 1 (March 1, 2012): 119–36. https://doi.org/10.1080/17524032.2011.644632.

Didier, Christelle, Weiwen Duan, Jean-Pierre Dupuy, David H. Guston, Yongmou Liu, José Antonio López Cerezo, Diane Michelfelder, et al. “Acknowledging AI’s Dark Side.” Science (New York, N.Y.) 349, no. 6252 (September 4, 2015): 1064–65. https://doi.org/10.1126/science.349.6252.1064-c.

Ding, Jeffrey. “Deciphering China’s AI Dream,” technical report, Future of Humanity Institute, 2018.

Douglas, Heather. Science, Policy, and the Value-Free Ideal. Pittsburgh, PA: University of Pittsburgh Press, 2009.

Editors, Artificial Intelligence and Law. “From the Editors:” Artificial Intelligence and Law 1, no. 1 (March 1, 1992): 1–2. https://doi.org/10.1007/BF00118476.

Editors, AI & Society. "Editorial." AI & SOCIETY 1, no. 1 (July 1, 1987): 3–4. https://doi.org/10.1007/BF01905884.

Elish, M. C. et al. "The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near Term," The AI Now Institute, 2016. https://ainowinstitute.org/AI_Now_2016_Report.pdf

Etzioni, Oren. "Most Experts Say AI Isn't as Much of a Threat as You Might Think." MIT Technology Review. 2016. Accessed March 9, 2018. https://www.technologyreview.com/s/602410/no-the-experts-dont-think-superintelligent-ai-is-a-threat-to-humanity/.

Evans, Benedict. "In What Senses Is Worrying about 'AI Morality'* Different to Worrying about 'Database Morality'? The Scenarios That Come up Seem Quite Similar: -'Computer Lets Bad People Do This Bad Thing' -'Computer Wrongly Tells Us to Do x' (*in Terms of How People Use AI, Not Skynet)." Tweet. @BenedictEvans (blog), March 14, 2018. https://twitter.com/BenedictEvans/status/973710360757731328.

Executive Office of the President. Preparing for the Future of Artificial Intelligence." Whitehouse.gov (archived), October 12, 2016. https://obamawhitehouse.archives.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence.

Farooque, Mahmud and Leak Kaplan. "Our Driverless Futures: Community Forums on Automated Mobility," website of the Consortium for Science, Policy, and Outcomes, accessed November 10, 2019. https://cspo.org/research/driverless-vehicles/

Fast, Ethan, and Eric Horvitz. "Long-Term Trends in the Public Perception of Artificial Intelligence." ArXiv:1609.04904 [Cs], September 15, 2016. http://arxiv.org/abs/1609.04904.

FATML. "Home : FAT ML." Accessed March 9, 2018. https://www.fatml.org/.

Felt, Ulrike, Clark Miller, and Laurel Smith-Doerr. The Handbook of Science and Technology Studies. Fourth edition. Cambridge, Massachusetts: MIT Press, 2017.

Ferguson, Andrew. "Rise of Big Data Policing." NYU Press. Accessed April 15, 2018. http://nyupress.org/books/9781479892822/.

Finn, Ed. What Algorithms Want. Cambridge, MA: MIT Press. 2017.

Fisher, Erik. “Lessons Learned from the Ethical, Legal and Social Implications Program (ELSI): Planning Societal Implications Research for the National Nanotechnology Program.” Technology in Society 27, no. 3 (August 2005): 321–28. https://doi.org/10.1016/j.techsoc.2005.04.006.

Fisher, Erik, Roop L. Mahajan, and Carl Mitcham. “Midstream Modulation of Technology: Governance From Within.” Bulletin of Science, Technology & Society 26, no. 6 (December 1, 2006): 485–96. https://doi.org/10.1177/0270467606295402.

Fisher, Erik, Michael O’Rourke, Robert Evans, Eric B. Kennedy, Michael E. Gorman, and Thomas P. Seager. “Mapping the Integrative Field: Taking Stock of Socio-Technical Collaborations.” Journal of Responsible Innovation 2, no. 1 (January 2, 2015): 39–61. https://doi.org/10.1080/23299460.2014.1001671.

Frickel, Scott, and Neil Gross. “A General Theory of Scientific/Intellectual Movements.” American Sociological Review 70, no. 2 (2005): 204–32.

Friedman, Benjamin M. The Moral Consequences of Economic Growth. New York, NY: Vintage Books, 2005.

Future of Life Institute. “AI Open Letter.” 2015. Future of Life Institute website. Accessed March 9, 2018. https://futureoflife.org/ai-open-letter/.

Future of Life Institute. “AI Principles.” Future of Life Institute website. 2017. Accessed March 9, 2018. https://futureoflife.org/ai-principles/.

Gallagher, Kelly Sims, Arnulf Grübler, Laura Kuhl, Gregory Nemet, and Charlie Wilson. “The Energy Technology Innovation System.” Annual Review of Environment and Resources 37, no. 1 (2012): 137–62. https://doi.org/10.1146/annurev-environ-060311-133915.

Gagne, Jean-Francois. “Global AI Talent Pool Report.” 2018. Accessed March 9, 2018. http://www.jfgagne.ai/talent/.

Gallagher, Elizabeth M. and Joanna J. Bryson. “Agent-Based Modeling,” Encyclopedia of Animal Cognition and Behavior, 2017. http://www.cs.bath.ac.uk/~jjb/ftp/GallagherBrysonABM-authorsfinal.pdf.

Geels, Frank W. “Technological Transitions as Evolutionary Reconfiguration Processes: A Multi-Level Perspective and a Case-Study.” Research Policy 31, no. 8–9 (December 2002): 1257–74. https://doi.org/10.1016/S0048-7333(02)00062-8.

Geraci, Robert. Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality. Oxford: Oxford University Press. Reprint edition, 2012.

Glaser, Charles L., and Chaim Kaufmann. “What Is the Offense-Defense Balance and Can We Measure It?” International Security 22, no. 4 (1998): 44. https://doi.org/10.2307/2539240.

Gong, Min, Robert J. Lempert, Andrew Parker, Lauren A. Mayer, Jordan R. Fischbach, Matthew Sisco, Zhimin Mao, David H. Krantz, and Howard Kunreuther. “Testing the Scenario Hypothesis.” RAND, 2017. https://www.rand.org/pubs/external_publications/EP67012.html.

Good, I. J. “Speculations Concerning the First Ultraintelligent Machine,” https://web.archive.org/web/20010527181244/http://www.aeiveos.com/~bradbury/Authors/Computing/Good-IJ/SCtFUM.html#CarterCF1963

Gottschall, Jonathan. The Storytelling Animal: How Stories Make Us Human. Boston: Mariner Books, 2013.

Grace, Katja et al., “2016 Expert Survey on Progress in AI,” 2016, AI Impacts blog. Accessed March 9, 2018. https://aiimpacts.org/2016-expert-survey-on-progress-in-ai/.

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. “When Will AI Exceed Human Performance? Evidence from AI Experts.” ArXiv:1705.08807 [Cs], May 24, 2017. http://arxiv.org/abs/1705.08807.

Grinbaum, Alexei, and Christopher Groves. “What Is ‘Responsible’ about Responsible Innovation? Understanding the Ethical Issues.” Responsible Innovation, Wiley Online Books, April 2, 2013. https://doi.org/10.1002/9781118551424.ch7.

Grunwald, Armin. “Responsible Innovation: Bringing Together Technology Assessment, Applied Ethics, and STS Research.” Responsible Innovation, 2011, 23.

———. “The Hermeneutic Side of Responsible Research and Innovation.” Wiley, 2016. https://www.wiley.com/en-us/The+Hermeneutic+Side+of+Responsible+Research+and+Innovation-p-9781786300850.

Guston, David H. Between Politics and Science: Assuring the Integrity and Productivity of Research. 1st US-1st Printing edition. Cambridge ; New York: Cambridge University Press, 2000.

———. "'Daddy, Can I Have a Puddle Gator?': Creativity, Anticipation, and Responsible Innovation." Responsible Innovation, Wiley Online Books, April 2, 2013. https://doi.org/10.1002/9781118551424.ch6.

———. "Understanding 'Anticipatory Governance.'" Social Studies of Science 44, no. 2 (April 1, 2014): 218–42. https://doi.org/10.1177/0306312713508669.

Guston, David H., and Daniel Sarewitz. "Real-Time Technology Assessment." Technology in Society, American Perspectives on Science and Technology Policy, 24, no. 1 (January 1, 2002): 93–109. https://doi.org/10.1016/S0160-791X(01)00047-1.

Guston, David H., Erik Fisher, Armin Grunwald, Richard Owen, Tsjalling Swierstra, and Simone van der Burg. "Responsible Innovation: Motivations for a New Journal." Journal of Responsible Innovation 1, no. 1 (January 2, 2014): 1–8. https://doi.org/10.1080/23299460.2014.885175.

Guston, David H. and Walter Valdivia. "Responsible Innovation: A Primer for Policymakers." Brookings Institution, 2015. https://www.brookings.edu/research/responsible-innovation-a-primer-for-policymakers/.

Hardt, Moritz, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning." ArXiv:1610.02413 [Cs], October 7, 2016. http://arxiv.org/abs/1610.02413.

Harremoës et al., eds."Late Lessons from Early Warnings: The Precautionary Principle 1896-2000." Publication. European Environment Agency. 2001. Accessed April 15, 2018. https://www.eea.europa.eu/publications/environmental_issue_report_2001_22.

Hanson, Robin. "The Age of Em: Work, Love, and Life When Robots Rule the Earth," Oxford University Press, 2016.

Hanson, Robin, and Yudkowsky, Eliezer. "The Hanson-Yudkowsky AI-Foom Debate." Machine Intelligence Research Institute. 2013. Accessed March 9, 2018. https://intelligence.org/ai-foom-debate/.

Hicks, Marie. “Programmed Inequality.” MIT Press. 2017. Accessed March 9, 2018. https://mitpress.mit.edu/books/programmed-inequality.

Hinton, Geoffrey. “AMA Geoffrey Hinton • r/MachineLearning.” Reddit, 2014. Accessed March 9, 2018. https://www.reddit.com/r/MachineLearning/comments/2lmo0l/ama_geoffrey_hinton/.

Hoop, Evelien de, Auke Pols, and Henny Romijn. “Limits to Responsible Innovation.” Journal of Responsible Innovation 3, no. 2 (May 3, 2016): 110–34. https://doi.org/10.1080/23299460.2016.1231396.

Horvitz, Thomas G. Dietterich, Eric J. “Rise of Concerns About AI: Reflections and Directions.” Accessed March 9, 2018. https://cacm.acm.org/magazines/2015/10/192386-rise-of-concerns-about-ai/abstract.

Howlett, Michael, M. Ramesh, and Anthony Perl. Studying Public Policy: Policy Cycles and Policy Subsystems. 3 edition. Ont. ; New York: OUP Canada, 2009.

Hughes, Thomas Parker. Networks of Power: Electrification in Western Society, 1880-1930. New Ed edition. Baltimore, Md.: Johns Hopkins University Press, 1993.

Hwang, Tim. “Computational Power and the Social Impact of Artificial Intelligence.” SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, March 23, 2018. https://papers.ssrn.com/abstract=3147971.

Illing, Sean. “How Worried Should We Be about Artificial Intelligence? I Asked 17 Experts.” Vox, March 8, 2017. https://www.vox.com/conversations/2017/3/8/14712286/artificial-intelligence-science-technology-robots-singularity-automation.

Islam, Riashat, Peter Henderson, Maziar Gomrokchi, and Doina Precup. “Reproducibility of Benchmarked Deep Reinforcement Learning Tasks for Continuous Control.” ArXiv:1708.04133 [Cs], August 10, 2017. http://arxiv.org/abs/1708.04133.

ITU. “Reality Check: ‘We Are Not Nearly as Close to Strong AI as Many Believe.’” ITU News (International Telecommunication Union), June 8, 2017. http://news.itu.int/reality-check-not-nearly-close-strong-ai-many-believe/.

Jasanoff, Sheila, J. Benjamin Hurlbut, and Krishanu Saha. “CRISPR Democracy: Gene Editing and the Need for Inclusive Deliberation,” Issues in Science and

Technology, Fall 2015. http://issues.org/32-1/crispr-democracy-gene-editing-and-the-need-for-inclusive-deliberation/.

Jasanoff, Sheila. The Ethics of Invention: Technology and the Human Future. New York: W. W. Norton & Company, 2016.

Jonas, Hans. "Toward a Philosophy of Technology." Hastings Center Report 9, no. 1 (February, 1979): 34–43. https://doi.org/10.2307/3561700.

Kahn, Jeremy. "Sky-High Salaries Are the Weapons in the AI Talent War." Bloomberg. February 13, 2018. https://www.bloomberg.com/news/articles/2018-02-13/in-the-war-for-ai-talent-sky-high-salaries-are-the-weapons.

Kaplan, Fred M. The Wizards of Armageddon. New York: Simon & Schuster Books, 1983.

Keeler, Lauren. "Quenching Our Thirst for Future Knowledge: Participatory Scenario Construction and Sustainable Water Governance in a Desert City" Dissertation, 2014. Accessed March 9, 2018. https://search.proquest.com/openview/81a8f761a20690e55ac7486cab0996a0/1?pq-origsite=gscholar&cbl=18750&diss=y.

Kerr, Anne, Rosemary L. Hill, and Christopher Till. "The Limits of Responsible Innovation: Exploring Care, Vulnerability and Precision Medicine." Technology in Society, Technology and the Good Society, 52 (February 1, 2018): 24–31. https://doi.org/10.1016/j.techsoc.2017.03.004.

Krakovna, Viktoriya. "Analysis: Clopen AI - Openness in Different Aspects of AI Development." Future of Life Institute, August 3, 2016. https://futureoflife.org/2016/08/03/op-ed-clopen-ai-openness-in-different-aspects-of-ai-development/.

———. "Is There a Tradeoff between Immediate and Longer-Term AI Safety Efforts?" Deep Safety (blog), January 27, 2018. https://vkrakovna.wordpress.com/2018/01/27/is-there-a-tradeoff-between-safety-concerns-about-current-and-future-ai-systems/.

Kroll, Joshua A., Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. "Accountable Algorithms." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, March 2, 2016. https://papers.ssrn.com/abstract=2765268.

Kurzweil, Raymond. The Singularity Is Near. London: Gerald Duckworth & Co Ltd, 2006.

Lawrence, Neil. “System Zero: What Kind of AI Have We Created?” April 12, 2015. Accessed April 15, 2018. http://inverseprobability.com/2015/12/04/what-kind-of-ai.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. “Deep Learning.” Nature 521, no. 7553 (May 27, 2015): 436–44. https://doi.org/10.1038/nature14539.

Lempert, Robert, Steven W. Popper, and Steven C. Bankes. Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis. Santa Monica, CA: RAND, 1999.

Leyton-Brown, Kevin and Yoav Shoham. Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations. Cambridge ; New York: Cambridge University Press, 2008.

Lipsey, Richard G., Kenneth I. Carlaw, and Clifford T. Bekar. Economic Transformations: General Purpose Technologies and Long-Term Economic Growth. Oxford, New York: Oxford University Press, 2005.

Lipton, Zachary C. “The Doctor Just Won’t Accept That!” ArXiv:1711.08037 [Stat], November 19, 2017. http://arxiv.org/abs/1711.08037.

Macnaghten, Phil, Richard Owen, J Stilgoe, Brian Wynne, Adalberto Azevedo, André de Campos, J Chilvers, et al. “Responsible Innovation across Borders: Tensions, Paradoxes and Possibilities.” Journal of Responsible Innovation 1 (May 12, 2014). https://doi.org/10.1080/23299460.2014.922249.

Manyika, James, Michael Chui, Mehdi Miremadi, Jacques Bughin, Katy George, Paul Willmott, and Martin Dewhurst. “Harnessing Automation for a Future That Works | McKinsey & Company.” Accessed March 9, 2018. https://www.mckinsey.com/global-themes/digital-disruption/harnessing-automation-for-a-future-that-works.

Martinez-Plumed, F., Shahar Avin, Miles Brundage, Allan Dafoe, Sean Ó hÉigeartaigh, and José Hernández-Orallo. “Accounting for the Neglected Dimensions of AI Progress,” arXiv preprint server, June 2, 2018. https://arxiv.org/abs/1806.00610

Masad, David, and Jacqueline Kazil. “Mesa: An Agent-Based Modeling Framework,” Python in Science Conference (SciPy), 2015.

Maschler, M. et al. Game Theory. Cambridge: Cambridge University Press, 2013.

Miller, Clark and Ira Bennett. "Thinking longer term about technology: Is there value in science fiction-inspired approaches to constructing futures?" Science and Public Policy, 35:8, 2008.

Moore, Kelly. Disrupting Science: Social Movements, American Scientists, and the Politics of the Military, 1945-1975. Princeton University Press. 2013. https://press.princeton.edu/titles/8545.html.

Morgan, M. Granger. "Use (and Abuse) of Expert Elicitation in Support of Decision Making for Public Policy." Proceedings of the National Academy of Sciences 111, no. 20 (May 20, 2014): 7176–84. https://doi.org/10.1073/pnas.1319946111.

———.. "Theory and Practice in Policy Analysis: Including Applications in Science and Technology." Cambridge University Press, August 2017. https://doi.org/10.1017/9781316882665.

Müller, Vincent C., and Bostrom, Nick. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." In Fundamental Issues of Artificial Intelligence, edited by Vincent C. Müller, 553–571. Springer, 2016.

Ng, Andrew, quoted by Williams, Chris,"AI Guru Ng: Fearing a Rise of Killer Robots Is like Worrying about Overpopulation on Mars." The Register, March 19, 2015. Accessed March 9, 2018. https://www.theregister.co.uk/2015/03/19/andrew_ng_baidu_ai/.

O'Neil, Cathy. "Weapons of Math Destruction, How Big Data Increases Inequality and Threatens Democracy." Penguin, 2016. https://www.penguin.co.uk/books/304513/weapons-of-math-destruction/.

Owen, Richard, Jack Stilgoe, Phil Macnaghten, Mike Gorman, Erik Fisher, and Dave Guston. "A Framework for Responsible Innovation." Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society 31 (2013): 27–50.

Parry, V. et al. "Principles of Robotics - EPSRC Website." 2011. Accessed April 15, 2018. https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/.

Parson, Edward A. "Opinion: Climate Policymakers and Assessments Must Get Serious about Climate Engineering." Proceedings of the National Academy of Sciences 114, no. 35 (August 29, 2017): 9227–30. https://doi.org/10.1073/pnas.1713456114.

Parson, Edward, Virginia Burkett, Karen Fisher-Vanden, David Keith, Linda Mearns, Hugh Pitcher, Cynthia Rosenzweig, and Mort Webster. “Global-Change Scenarios: Their Development and Use.” Other Publications, January 1, 2007. https://repository.law.umich.edu/other/35.

Partnership on AI. “Home.” Partnership on Artificial Intelligence to Benefit People and Society. Accessed April 15, 2018. https://www.partnershiponai.org/.

Piketty, Thomas, and Goldhammer, Arthur (translator). Capital in the Twenty-First Century. Cambridge Massachusetts: Harvard University Press, 2014.

Radford, Alec, Jeff Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. “Better Language Models and their Implications,” OpenAI blog post, February 2019. https://openai.com/blog/better-language-models/

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. “Language Models are Unsupervised Multitask Learners,” OpenAI technical paper, February 2019.

Ramírez, Rafael, and Cynthia Selin. “Plausibility and Probability in Scenario Planning.” Foresight 16, no. 1 (March 4, 2014): 54–74. https://doi.org/10.1108/FS-08-2012-0061.

Ramírez, Rafael, and Angela Wilkinson. Strategic Reframing: The Oxford Scenario Planning Approach. Oxford, New York: Oxford University Press, 2016.

Randles, Sally, Jan Youtie, David Guston, Barbara Harthorn, Chris Newfield, Philip Shapira, Fern Wickson, Arie Rip, Rene Schomberg, and Nick Pidgeon. A Transatlantic Conversation on Responsible Innovation and Responsible Governance, 2011.

Rip, Arie. “The Past and Future of RRI.” Life Sciences, Society and Policy 10 (November 6, 2014): 17. https://doi.org/10.1186/s40504-014-0017-4.

Royal Society. “Machine Learning Requires Careful Stewardship Says Royal Society.” Royal Society website, April 2017. Accessed March 9, 2018. https://royalsociety.org/news/2017/04/machine-learning-requires-careful-stewardship-says-royal-society/.

Russell, Stuart and Allan Dafoe. “Yes, the Experts Are Worried about the Existential Risk of Artificial Intelligence.” MIT Technology Review. 2016. Accessed March 9, 2018. https://www.technologyreview.com/s/602776/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/.

Russell, Stuart, and Peter Norvig. Artificial Intelligence: A Modern Approach. 3 edition. Upper Saddle River: Pearson, 2009.

Sachs, Jeffrey. Common Wealth: Economics for a Crowded Planet. First Printing edition. London: Allen Lane, 2008.

Sadowski, Jathan. "Office of Technology Assessment: History, Implementation, and Participatory Critique." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, January 1, 2015. https://papers.ssrn.com/abstract=2555364.

Sarewitz, Daniel, and Richard Nelson. "Three Rules for Technological Fixes." Comments and Opinion. Nature, December 17, 2008. https://doi.org/10.1038/456871a.

Sarewitz, Daniel (ed.), Miles Brundage, Travis Doom, Eric B. Kennedy, Frank N. Laird, Jason O'Leary, Heather M. Ross, Aubrey Wigner, and Jen Fuller. The Rightful Place of Science: Government & Energy Innovation. Consortium for Science, Policy & Outcomes, 2014.

Scharre, Paul. "Army of None: Autonomous Weapons and the Future of War." W. W. Norton & Company." 2018. Accessed April 15, 2018. http://books.wwnorton.com/books/978-0-393-60898-4/.

Schot, Johan, and Arie Rip. "The Past and Future of Constructive Technology Assessment." Technological Forecasting and Social Change, Technology Assessment: The End of OTA, 54, no. 2 (February 1, 1997): 251–68. https://doi.org/10.1016/S0040-1625(96)00180-1.

Scott, Andrew C, José R Solórzano, Jonathan D Moyer, and Barry B Hughes. "Modeling Artificial Intelligence and Exploring Its Impact," Pardee Center for International Futures, May 2017. https://pardee.du.edu/sites/default/files/ArtificialIntelligenceIntegratedPaper_V6_clean.pdf

Selin, Cynthia. "Professional Dreamers: The Future In The Past Of Scenario Planning." Scenarios For Success, Wiley Online Books, October 16, 2015. https://doi.org/10.1002/9781119208136.ch2.

Simonite, Tom. "For Superpowers, Artificial Intelligence Fuels New Global Arms Race," Wired, September 8, 2017. https://www.wired.com/story/for-superpowers-artificial-intelligence-fuels-new-global-arms-race/.

Sipser, Michael. Introduction to the Theory of Computation. 3 edition. Boston, MA: Course Technology, 2012.

Shoham, Yoav, Raymond Perrault, Erik Brynjolfsson, Jack Clark, James Mankiya, Juan Carlos Niebles, Terah Lyons, John Etchemendy, Barbara Grosz, and Zoe Bauer. "AI Index 2018," Stanford AI Index project, available online at http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf

Slottje, P., J. P. van der Sluijs, and A. B. Knol. "Expert elicitation: methodological suggestions for its use in environmental health impact assessments." Book, 2008. http://dspace.library.uu.nl/handle/1874/32938.

Slusallek, Philipp. "Artificial Intelligence and Digital Reality: Do We Need a CERN for AI?" The Forum Network, hosted by the OECD, January 8, 2018. https://www.oecd-forum.org/users/71431-philipp-slusallek/posts/28452-artificial-intelligence-and-digital-reality-do-we-need-a-cern-for-ai.

Solaiman, Irene, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. "Release Strategies and the Social Impacts of Language Models," arXiv preprint server, 2019. https://arxiv.org/abs/1908.09203

Spruit, Shannon, Gordon Hoople, and David Rolfe. "Just a Cog in the Machine? The Individual Responsibility of Researchers in Nanotechnology to Collectivize," Science and Engineering Ethics, 2016, 22: 871-887.

Stilgoe, Jack. Experiment Earth. First edition. London, New York: Routledge, 2016.

Stilgoe, Jack, Richard Owen, and Phil Macnaghten. "Developing a Framework for Responsible Innovation." Research Policy 42, no. 9 (November 1, 2013): 1568–80. https://doi.org/10.1016/j.respol.2013.05.008.

Stirling, Andy. "'Opening Up' and 'Closing Down': Power, Participation, and Pluralism in the Social Appraisal of Technology." Science, Technology, & Human Values 33, no. 2 (March 1, 2008): 262–94. https://doi.org/10.1177/0162243907311265.

Stone, Peter et al. "2016 Report | One Hundred Year Study on Artificial Intelligence (AI100)." Accessed March 9, 2018. https://ai100.stanford.edu/2016-report.

Sutskever, Ilya, and Dario Amodei. "Protecting Against AI's Existential Threat." Wall Street Journal, October 18, 2017, sec. Life. https://www.wsj.com/articles/protecting-against-ais-existential-threat-1508332313.

Theodorou, Andreas, Bryn Bandt-Law, and Joanna Bryson. "The Sustainability Game: AI Technology as an Intervention for Public Understanding of

Cooperative Investment," IEEE Conference on Games (CoG), August 2019. http://www.cs.bath.ac.uk/~jjb/ftp/TheodorouBandtLawBrysonCoG19.pdf

Tomblin, David, Zachary Pirtle, Mahmud Farooque, David Sittenfeld, Erin Mahoney, Rick Worthington, Gretchen Gano, et al. "Integrating Public Deliberation into Engineering Systems: Participatory Technology Assessment of NASA's Asteroid Redirect Mission." Astropolitics 15, no. 2 (May 4, 2017): 141–66. https://doi.org/10.1080/14777622.2017.1340823.

Tucker, Philip. Innovation, Dual Use, and Security: Managing the Risks of Emerging Biological and Chemical Technologies. Cambridge, MA: MIT Press, 2012.

Turing, Alan. "Lecture to the London Mathematical Society on 20 February 1947," 1947. Accessed April 15, 2018. https://www.vordenker.de/downloads/turing-vorlesung.pdf.

US Department of Energy. The Quadrennial Energy Review (QER), US Department of Energy website 2017. Accessed April 15, 2018. https://www.energy.gov/policy/initiatives/quadrennial-energy-review-qer.

Usher, Will, and Neil Strachan. "An Expert Elicitation of Climate, Energy and Economic Uncertainties." Energy Policy 61 (October 1, 2013): 811–21. https://doi.org/10.1016/j.enpol.2013.06.110.

van der Heijden, Kees. Scenarios: The Art of Strategic Conversation (second edition), 2005. Wiley.

Von Schomberg, Rene. "Towards Responsible Research and Innovation in the Information and Communication Technologies and Security Technologies Fields." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, November 13, 2011. https://papers.ssrn.com/abstract=2436399.

Weizenbaum, Joseph. Computer Power and Human Reason: From Judgment to Calculation. 1st edition. San Francisco: W H Freeman & Co, 1976.

Weld, Daniel, and Oren Etzioni. "The First Law of Robotics (A Call to Arms)," AAAI-94, 1994. https://www.aaai.org/Papers/AAAI/1994/AAAI94-160.pdf

Weller, Adrian. "Challenges for Transparency." arXiv preprint server, July 29, 2017. http://arxiv.org/abs/1708.01870.

Weninger, Samim. "Sensual Machines." Samim (blog), June 29, 2015. https://medium.com/@samim/sensual-machines-82858b32a4e5.

The White House. “The Administration’s Report on the Future of Artificial Intelligence.” whitehouse.gov, October 12, 2016. https://obamawhitehouse.archives.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence.

Wiener, Norbert. “Some Moral and Technical Consequences of Automation,” Science, Volume 131, Issue 3410, pp. 1355-1358, May 6, 1960.

———. “God & Golem, Inc.” The MIT Press. 1964. https://mitpress.mit.edu/books/god-golem-inc.

Wilsdon, James, and Rebecca Willis. “See-through Science: Why Public Engagement Needs to Move Upstream.” Reports and working papers, September 2004. http://www.demos.co.uk/publications/paddlingupstream.

Wikipedia, “Overton Window.” Wikipedia authors, Accessed April 5, 2018. https://en.wikipedia.org/w/index.php?title=Overton_window&oldid=834389918 .

Wilensky, Uri and William Rand. An Introduction to Agent-Based Modeling. The MIT Press. 2015. https://mitpress.mit.edu/books/introduction-agent-based-modeling.

WiML. “Women in Machine Learning.” Accessed March 9, 2018. http://wimlworkshop.org/.

Winner, Langdon. The Whale and the Reactor: A Search for Limits in an Age of High Technology. Chicago: University of Chicago Press, 1986.

Wong, Pak-Hang. “Responsible Innovation for Decent Nonliberal Peoples: A Dilemma?” Journal of Responsible Innovation 3, no. 2 (May 3, 2016): 154–68. https://doi.org/10.1080/23299460.2016.1216709.

Wu, Tim. “Opinion | Please Prove You’re Not a Robot.” The New York Times, July 15, 2017, sec. Opinion. https://www.nytimes.com/2017/07/15/opinion/sunday/please-prove-youre-not-a-robot.html.

Yudkowsky, Eliezer. “Intelligence Explosion Microeconomics.” Machine Intelligence Research Institute, Accessed Online October 23 (2013): 2015.

APPENDIX A

SURVEY QUESTIONS

1. By what year do you expect a 50% chance of high-level machine intelligence having been created? "High-level machine intelligence” (HLMI) is considered to be achieved when unaided machines can accomplish most tasks better and more cheaply than human workers. [open-ended]

2. Would you characterize the field of AI today as being open or closed? By "open" we mean a field in which research ideas and results are widely disseminated quickly after their development, and by "closed" we mean a field in which research ideas and results are closely guarded secrets. [multiple choice--completely open, mostly open, an even mixture of open vs. closed, mostly closed, completely closed]

3. How confident are you in your last answer, on a scale from 0 to 100, with 0 meaning that you might as well have chosen randomly, and 100 meaning that you are absolutely sure? [0-100--repeated after each of the next 12 questions]

4. Around the time that high-level machine intelligence is developed, do you expect that the AI field will be open or closed? By "open" we refer to a field in which research ideas and results are widely disseminated quickly after their development, and by "closed" we refer to a field in which research ideas and results are closely guarded secrets. [multiple choice--completely open, mostly open, an even mixture of open vs. closed, mostly closed, completely closed]

5. Around the time that high-level machine intelligence is developed, do you think that the field of AI should be open or closed? By "open" we refer to a field in which research ideas and results are widely disseminated quickly after their development, and by "closed" we refer to a field in which research ideas and results are closely guarded secrets. [multiple choice--completely open, mostly open, an even mixture of open vs. closed, mostly closed, completely closed]

6. Would you characterize AI capabilities today as being distributed or concentrated? By "distributed" we mean that everyone is able to deploy the most powerful existing AI techniques, and by "concentrated" we mean that only a very small number of people is able to deploy the most powerful existing AI techniques. [multiple choice--totally distributed, mostly distributed, an even mixture of distributed vs. concentrated, mostly concentrated, totally concentrated]

7. After high-level machine intelligence (HLMI) is developed, do you expect AI capabilities to be distributed or concentrated? By "distributed" we mean that everyone is able to deploy the most powerful existing AI techniques, and by "concentrated" we mean

that only a very small number of people is able to deploy the most powerful existing AI techniques. [multiple choice--totally distributed, mostly distributed, an even mixture of distributed vs. concentrated, mostly concentrated, totally concentrated]

8. After high-level machine intelligence (HLMI) is developed, do you think that AI capabilities should be distributed or concentrated? By "distributed" we mean that everyone is able to deploy the most powerful existing AI techniques, and by "concentrated" we mean that only a very small number of people is able to deploy the most powerful existing AI techniques. [multiple choice--totally distributed, mostly distributed, an even mixture of distributed vs. concentrated, mostly concentrated, totally concentrated]

9. How involved do you think the United States government is today in AI research, development, and deployment? [multiple choice--very involved, somewhat involved, not at all involved]

10. How involved do you think the United States government will be in AI research, development, and deployment around the time that HLMI is developed? [multiple choice--very involved, somewhat involved, not at all involved]

11. How involved do you think the United States government should be in AI research, development, and deployment around the time that HLMI is developed? [multiple choice--very involved, somewhat involved, not at all involved]

12. How involved do you think the Chinese government is today in AI research, development, and deployment? [multiple choice--very involved, somewhat involved, not at all involved]

13. How involved do you think the Chinese government will be in AI research, development, and deployment around the time that HLMI is developed? [multiple choice--very involved, somewhat involved, not at all involved]

14. How involved do you think the Chinese government should be in AI research, development, and deployment around the time that HLMI is developed? [multiple choice--very involved, somewhat involved, not at all involved]

15. Which do you think is most likely to cause harm: AI being used deliberately for malicious purposes, or AI causing harm that was unintended by humans? [multiple choice--very involved, somewhat involved, not at all involved]

After questions 1-15, a follow-up question was asked after each, which is reproduced only once here for concision: “How confident are you in your last answer, on a scale from 0 to 100, with 0 meaning that you might as well have chosen randomly, and 100 meaning that you are absolutely sure?” [0-100]

16. What is your gender? [female/male/other]

17. What is your age? [17 or younger, 18-20, 21-29, 30-39, 40-49, 50-59, 60 or older]

18. What is the highest level of school that you have completed? [primary school, some high school but no diploma, some college but no degree, 2-year college degree, 4-year college degree, graduate-level degree, none of the above]

19. Which race/ethnicity best describes you? (Please choose only one.) [Asian/Pacific Islander, Black/African/Caribbean, Hispanic, White/Caucasian, First Nation/American Indian/Indigenous, Mixed/Multiple ethnic groups, Other]

20. How long have you been studying the science/technology of AI (in years)? [open ended]

21. How long have you been studying the policy/strategy aspects of AI (in years)? [open ended]

22. What is your name? [open ended]

Remaining questions are only from the post-workshop survey. Otherwise, the questions were identical pre and post.

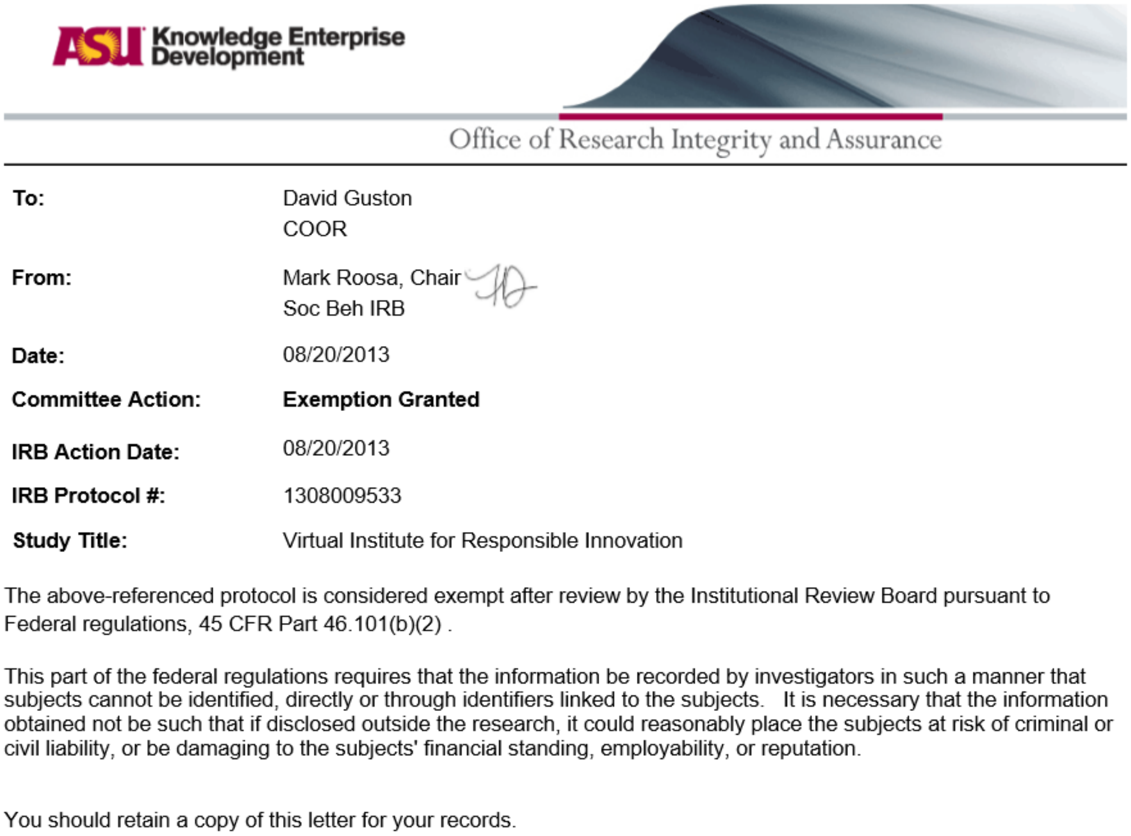
23. Did your expectations regarding the future of AI change in any noticeable way as a result of participation in the scenario planning workshop? If so, please briefly describe this below, and leave the box blank otherwise. [open ended]

24. Did your preferences regarding the future of AI change in any noticeable way as a result of participation in the scenario planning workshop? If so, please briefly describe this below, and leave the box blank otherwise. [open ended]

25. Please use the box below if you would like to provide any additional feedback on the structure, content, and/or usefulness of the workshop. [open ended]

APPENDIX B

INSTITUTIONAL REVIEW BOARD DOCUMENTATION



ASU Knowledge Enterprise Development

Office of Research Integrity and Assurance

To:	David Guston COOR
From:	Mark Roosa, Chair Soc Beh IRB
Date:	08/20/2013
Committee Action:	**Exemption Granted**
IRB Action Date:	08/20/2013
IRB Protocol #:	1308009533
Study Title:	Virtual Institute for Responsible Innovation

The above-referenced protocol is considered exempt after review by the Institutional Review Board pursuant to Federal regulations, 45 CFR Part 46.101(b)(2) .

This part of the federal regulations requires that the information be recorded by investigators in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects. It is necessary that the information obtained not be such that if disclosed outside the research, it could reasonably place the subjects at risk of criminal or civil liability, or be damaging to the subjects' financial standing, employability, or reputation.

You should retain a copy of this letter for your records.