AI Governance: A Research Agenda

Allan Dafoe

Governance of AI Program, Future of Humanity Institute, University of Oxford[1]

First draft July 2017
v1.0 August 27 2018

Abstract

Artificial intelligence (AI) is a potent general purpose technology. Future progress could be rapid, and experts expect that superhuman capabilities in strategic domains will be achieved in the coming four decades. The opportunities are tremendous, including advances in medicine and health, transportation, energy, education, science, economic growth, and environmental sustainability. The risks, however, are also substantial and plausibly pose extreme governance challenges. These include labor displacement, inequality, an oligopolistic global market structure, reinforced totalitarianism, shifts and volatility in national power, strategic instability, and an AI race that sacrifices safety and other values. The consequences are plausibly of a magnitude and on a timescale to dwarf other global concerns, leaders of governments and firms are asking for policy guidance, and yet scholarly attention to the AI revolution remains negligible. Research is thus urgently needed on the AI governance problem: the problem of devising global norms, policies, and institutions to best ensure the beneficial development and use of advanced AI. This report outlines an agenda for this research.

[1] This document received input from many contributors. The text was primarily written by Allan Dafoe. Several portions received substantial input from other individuals, most affiliated with the Governance of AI Program, noted where for each portion. This work draws from the body of thinking and insight in the community of scholars and scientists thinking about these issues. In particular, for comments, conversations, and related work this document benefits from Miles Brundage, Owen Cotton-Barratt, Richard Danzig, Daniel Dewey, Owain Evans, Paul de Font-Reaulx, Genevieve Fried, Ben Garfinkel, Katja Grace, Rox Heston, Geoffrey Irving, Charlotte Jander, Jade Leung, Chris Olah, Catherine Olsson, Cullen O'Keefe, Andrew Snyder-Beattie, Ken Schultz, Claudia Shi, Duncan Snidal, Nate Soares, Jaan Tallinn, Robert Trager, Helen Toner, Brian Tse, Eliezer Yudkowsky, Baobao Zhang, and especially Dario Amodei, Nick Bostrom, Owain Evans, Carrick Flynn, Rose Hadshar, Holden Karnofsky, Jan Leike, Matthijs Maas, Luke Muehlhauser, Toby Ord, Michael Page, Carl Shulman, and Remco Zwetsloot.

Preface

This document is meant to help introduce and orient researchers to the space of plausibly important problems in AI governance. It offers a framing of the overall problem, an attempt to be (at least superficially) comprehensive in posing questions that could be pivotal, and references to published articles relevant to these questions. Some disclaimers are in order.

(1) **Focus on extreme risks**: This document focuses on extreme risks from advanced AI. This is not necessarily meant to imply that advanced AI poses a high probability of extreme dangers, or that the most important risks are the extreme ones. This document focuses on risks more than opportunities for several reasons, including that they are often more time-sensitive to anticipate, they are often less likely to be addressed by the market, and many people are loss averse leading to substantial welfare gains from preventing risks. This document focuses on extreme risks more than moderate and small risks because it is addressed to a community of scholars, policymakers, and philanthropists who prioritize addressing extreme stakes, such as many in the Effective Altruism community and at our collaborating institutions.[2]

(2) **Not sufficient:** This document is **an** introduction, not **the** introduction. To calibrate expectations, consider that it takes years of reading, thinking, conversations, and trial and error for aspiring researchers in mature fields to find a way to make a contribution. For a new field like AI Governance, you should expect to at least take several months.

(3) **Neither parsimonious nor comprehensive**: There is a tradeoff between parsimony and comprehensiveness. This document may be unsatisfying on both fronts. It is not parsimonious: reading can feel like a "firehose" of questions and ideas. Nor is it close to comprehensive or detailed: some sentences summarize a large body of thought, there are many connections that are not made explicit. If you want more parsimony, look at the table of

[2] For an introduction to this perspective, see Karnofsky, Holden. "Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity." Open Philanthropy Project, 2016.;
Dewey, Daniel. "Potential Risks from Advanced AI." In *Effective Altruism Handbook*, 2nd edition, 2018.

contents, focus on bolded terms and topic sentences. If you want more comprehensiveness, follow the references in a given section.

(4) **Long and dense**: In aiming for (superficial) comprehensiveness, this document is long and dense. You should not expect to digest it in one sitting. You may want to skim most of it.

(5) **No single focus**: You should not expect this document to give you strong advice about what work to prioritize. This is a deliberate choice. For any given individual I can give recommendations of what work should be prioritized. For the community of researchers, I believe there is a vast range of work that should be done and that individuals should specialize according to their comparative advantage, interest, and insight. More on this below.

(6) **Not authoritative**: This document aims for (superficial) comprehensiveness, and so covers topics beyond my expertise. These topics may be discussed in less detail than their importance and tractability warrants. I encourage relevant experts to write short reviews for topics that are currently neglected.

This space is rapidly evolving. This document will be updated to reference new work, and to reflect the changing research landscape. Comments, suggestions, references, and questions are very welcome, as they will help improve later versions of this document. Please email them (to info@governance.ai) with the subject line 'Research agenda feedback'.

Contents

Introduction

Artificial intelligence[3] is likely to become superhuman at most important tasks within this century. This would pose tremendous opportunities and risks for humanity. Further, AI experts foresee a non-trivial probability that the next decade (~10%) or two (~25%)[4] could see AI capabilities emerge that could radically transform welfare, wealth, or power, to the extent of the nuclear revolution or the industrial revolution. These possibilities are strikingly neglected, in part because they involve massive global and intergenerational externalities. There is thus a high leverage opportunity to address what may be the most important global issue of the 21st century. Seeking to do this, the field of **AI Governance** studies *how humanity can best navigate the transition to advanced AI systems*,[5] focusing on the political, economic, military, governance, and ethical dimensions. This document provides an overview of this research landscape.

AI governance is often paired with AI safety.[6] Both have the goal of helping humanity develop beneficial AI. AI safety focuses on the technical questions of how AI is built; AI governance focuses on the political contexts in which AI is built and used. Specifically, AI governance seeks to maximize the odds that people building and using advanced AI have the goal,

[3] Defined simply as the development of machines capable of sophisticated (intelligent) information processing. Compare also the definition by Nils Nilsson: "Artificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment." Nilsson, Nils J. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge; New York: Cambridge University Press, 2010. For a survey of definitions of 'intelligence', see Legg, Shane, and Marcus Hutter. "A Collection of Definitions of Intelligence." *ArXiv:0706.3639 [Cs]*, June 25, 2007. http://arxiv.org/abs/0706.3639; for different definitions of 'AI' used within the field, see Russell, S.J., and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Prentice Hall, 2009 (3rd edition). http://www.cin.ufpe.br/~tfl2/artificial-intelligence-modern-approach.9780131038059.25368.pdf, p. 5.

[4] These probabilities refer to the median respondent's beliefs about when we will see AI that is better than all humans at all tasks. For more on AI experts beliefs, see: 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; see also Grace, Katja. "2016 Expert Survey on Progress in AI." *AI Impacts*, December 14, 2016. https://aiimpacts.org/2016-expert-survey-on-progress-in-ai/. Note that these forecasts should not be regarded as especially reliable, given that the respondents are not known to be calibrated in their statements of probabilities, the respondents are not experts at forecasting nor are experts at macro-developments in AI; nevertheless, it provides a perspective on timelines. Interestingly, informally I would say the median of other more careful approaches yields a similar timeline.

[5] "Advanced AI" gestures towards systems substantially more capable (and dangerous) than existing (2018) systems, without necessarily invoking specific generality capabilities or otherwise as implied by concepts such as "Artificial General Intelligence" ("AGI").

[6] 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; Russell, Stuart, Daniel Dewey, and Max Tegmark. "Research Priorities for Robust and Beneficial Artificial Intelligence." *Future of Life Institute - AI Magazine*, 2015. https://futureoflife.org/data/documents/research_priorities.pdf?x90991; Everitt, Tom, Gary Lea, and Marcus Hutter. "AGI Safety Literature Review." *ArXiv:1805.01109 [Cs]*, May 3, 2018. http://arxiv.org/abs/1805.01109; Metz, Cade. "Teaching A.I. Systems to Behave Themselves." *The New York Times*, August 13, 2017, sec. Technology. https://www.nytimes.com/2017/08/13/technology/artificial-intelligence-safety-training.html.

motivation, worldview, time, training, resources, support, and organizational home necessary to do so for the benefit of humanity.

To motivate this work it can be helpful to consider an urgent, though not implausible, hypothetical scenario. Suppose that in one year's time a leading AI lab perceives that profound progress may be on the horizon. It concludes that given a big push, 6 to 24 months it is likely to develop techniques that would achieve breakthroughs of superhuman capabilities in strategic domains; these domains might include lie detection, social-network mapping and manipulation, cyber-operations, signals and imagery intelligence, strategy, bargaining or persuasion, engineering, science, and potentially AI research itself. Despite our knowledge (in this scenario) that these technical breakthroughs are likely, we would have uncertainty about the details, including: which transformative capabilities will come first and how they will work; how successive (small) capabilities may interact to become jointly transformative; how to build advanced AI in a safe way, and how difficult it will be to do so; what deployment plans and governance regimes will be most likely to lead to globally beneficial outcomes.

The AI governance problem is the problem of preparing for this scenario, as well as all other high-stakes implications of advanced AI. The task is substantial. What do we need to know and do in order to maximize the chances of the world safely navigating this transition? What advice can we give to AI labs, governments, NGOs, and publics, now and at key moments in the future? What international arrangements will we need--what vision, plan, technologies, protocols, organizations--to avoid firms and countries dangerously racing for short-sighted advantage? What will we need to know and arrange in order to elicit and integrate people's values, to deliberate with wisdom, and to assure groups so that they do not act out of fear?

The potential upsides of AI are tremendous. There is little that advanced intelligence couldn't help us with. Advanced AI could play a crucial role solving existing global problems, from climate change to international conflict. Advanced AI could help us dramatically improve health, happiness, wealth, sustainability, science, and self-understanding.[7]

[7] Bostrom, Nick, Allan Dafoe, and Carrick Flynn. "Public Policy and Superintelligent AI: A Vector Field Approach." Future of Humanity Institute, 2018. http://www.nickbostrom.com/papers/aipolicy.pdf.

The potential downsides, however, are also extreme. Let us consider four sources of catastrophic risk.

- (1) **Robust totalitarianism** could be enabled by advanced lie detection, social manipulation, autonomous weapons, and ubiquitous physical sensors and digital footprints. Power and control could radically shift away from publics, towards elites and especially leaders, making democratic regimes vulnerable to totalitarian backsliding, capture, and consolidation.
- (2) **Preventive, inadvertent, or unmanageable great-power (nuclear) war**. Advanced AI could give rise to extreme first-strike advantages, power shifts, or novel destructive capabilities, each of which could tempt a great power to initiate a preventive war. Advanced AI could make crisis dynamics more complex, unpredictable, and escalate faster than humans could manage, increasing the risk of inadvertent war.
- (3) Advanced AI systems could be built that **are not fully aligned with human values**, leading to human extinction or other permanent loss in value.[8] This risk is likely much greater if labs and countries are racing to develop and deploy advanced AI,[9] as researching and implementing AI safety measures is plausibly time and resource intensive.
- (4) Finally, even if we escape the previous three acute risks, we could experience systematic **value erosion from competition**, in which each actor repeatedly confronts a steep trade-off between pursuing their final values or pursuing the instrumental goal of adapting to the competition so as to have more power and wealth.[10]

[8] Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014. Yudkowsky, Eliezer. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Cirkovic. New York: Oxford University Press, 2008, pp. 308–45. See also the reading syllabus by Bruce Schneier. "Resources on Existential Risk - for: Catastrophic Risk: Technologies and Policies." Berkman Center for Internet and Society, Harvard University, 2015. https://futureoflife.org/data/documents/Existential%20Risk%20Resources%20(2015-08-24).pdf?x70892.

[9] Armstrong, Stuart, Nick Bostrom, and Carl Shulman. "Racing to the Precipice: A Model of Artificial Intelligence Development." Technical Report. Future of Humanity Institute, 2013. https://www.fhi.ox.ac.uk/wp-content/uploads/Racing-to-the-precipice-a-model-of-artificial-intelligence-development.pdf.

[10] Thanks to Daniel Dewey for suggesting this clear statement of the risks.

These risks can be understood as negative externalities: harms from socio-technical developments that impact individuals other than those responsible for the developments. These externalities are especially challenging to manage as they may be extreme in magnitude (extinction), complex and hard to predict, and they will spill across borders and generations. Building the right institutions and political arrangements is plausibly close to a necessary and sufficient condition to adequately address these risks. With the right institutions and political arrangements these risks can be radically reduced and plausibly eliminated. Without them, it may be that nothing short of a technical miracle will be sufficient to safely navigate the transition.

Transformative AI

Stepping back from this scenario, our research on AI governance strives to study AI's most transformative potential capabilities, dynamics, and impacts. The stakes could be extreme: absent an interruption in development, AI this century is likely to be sufficiently transformative as to "precipitate a transition comparable to (or more significant than) the agricultural or industrial revolution."[11] Given our current uncertainty[12] about what capabilities will have the greatest impacts, this document directs attention to a broad range of potentially transformative capabilities and dynamics.

The emergence of transformative innovations may be anticipated, gradual, and sequential, or they may emerge suddenly from a "discontinuous" jump in a broad range of capabilities, possibly due to recursive self-improvement, general intelligence, or abundant untapped computational power.

[11] Karnofsky, Holden. "Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity." Open Philanthropy Project, 2016. http://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity; Muehlhauser, Luke. "How Big a Deal Was the Industrial Revolution?" 2017. http://lukemuehlhauser.com/industrial-revolution/.

[12] As an analogy for the difficulty we may have in perceiving a transformative capability, it took about 10 years from Fleming's discovery of penicillin to the production of a compelling proof-of-concept and recognition by a major funder (Rockefeller) that this was worth seriously investing in. Bud, Robert. *Penicillin: Triumph and Tragedy*. Oxford: Oxford University Press, 2008, pp. 23–34.

Likewise with airplanes: in 1901, two years before building the first heavier-than-air plane, Wilbur Wright said to his brother "that men would not fly for fifty years." McCullough, David. *The Wright Brothers*. Simon & Schuster. p. 208; also quoted in Yudkowsky, Eliezer. "There's No Fire Alarm for Artificial General Intelligence." Machine Intelligence Research Institute, October 13, 2017. https://intelligence.org/2017/10/13/fire-alarm/.

What are some ways that AI could be transformative? AI could vastly increase wealth, health, and well-being. AI could transform work by radically altering employment prospects[13] or job security. It could increase economic inequality, domestically and globally. It could provide new tools of state repression and control, empowering authoritarian governments; it could also enable new forms of effective democratic decision-making and accountability, empowering democracy. It could transform international political economy (IPE); for example, AI is increasingly perceived as a strategic industry, activating massive industrial policy to support national AI champions and assets.

AI could transform international security by altering key strategic parameters, such as the security of nuclear retaliation,[14] the offense-defense balance, the stability of crisis escalation, the efficiency of negotiations, the viability of mutual privacy preserving surveillance, and the volatility and predictability of the future balance of power. It could enable new operational and strategic capabilities, such as in mass-persuasion, cyber-operations, command and control, intelligence, air combat, subsea combat, materials science, engineering, and science. These advantages may come in sufficient strength or combinations to radically transform power. Even the mere perception by governments and publics of such military (or economic) potential could lead to a radical break from the current technology and world order: shifting AI leadership to governments, giving rise to a massively funded AI race and potentially the securitization of AI development and capabilities.[15] This could undermine the liberal world economic order.[16] The intensity from a race dynamic could lead to catastrophic

[13] Aghion, Philippe, Benjamin Jones, and Charles Jones. "Artificial Intelligence and Economic Growth." Cambridge, MA: Stanford Institute for Economic Policy Research (SIEPR), October 2017. https://pdfs.semanticscholar.org/b0f0/989edd61ffa192c2a54e8edded9b84781719.pdf; Korinek, Anton, and Joseph E. Stiglitz. "Artificial intelligence and its implications for income distribution and unemployment." No. w24174, National Bureau of Economic Research, 2017. https://www8.gsb.columbia.edu/faculty/jstiglitz/sites/jstiglitz/files/w24174.pdf.

[14] Geist, Edward, and Andrew J Lohn. "How Might Artificial Intelligence Affect the Risk of Nuclear War?" RAND, 2018. https://www.rand.org/pubs/perspectives/PE296.html; Lieber, Keir A., and Daryl G. Press. "The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence." International Security 41, no. 4 (April 2017): 9–49. https://doi.org/10.1162/ISEC_a_00273; for work on the general interface of nuclear weapons with cybersecurity, see Futter, Andrew. *Hacking the Bomb: Cyber Threats and Nuclear Weapons.* Washington, DC: Georgetown University Press, 2018.

[15] On the deleterious effects of framing AI development as a 'race', see Cave, Stephen, and Seán S. Ó hÉigeartaigh. "An AI Race for Strategic Advantage: Rhetoric and Risks." In *AAAI / ACM Conference on Artificial Intelligence, Ethics and Society*, 2018. http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_163.pdf.

[16] Danzig, Richard, ed. "An Irresistible Force Meets a Moveable Object: The Technology Tsunami and the Liberal World Order." *Lawfare Research Paper Series* 5, no. 1 (August 28, 2017). https://assets.documentcloud.org/documents/3982439/Danzig-LRPS1.pdf.

corner-cutting from hurried development and deployment of (unsafe) advanced AI systems.[17] This danger poses an extreme urgency, and opportunity, for global cooperation.

More long-term and abstractly, the emergence of machine **superintelligence** (AI that is vastly better than humans at all important tasks) would enable revolutionary changes, more profound than the agricultural or industrial revolutions. Superintelligence would involve the bringing into existence of a new form of highly-capable intelligence, and plausibly one that could rapidly self-improve. Superintelligence offers tremendous opportunities, such as the radical reduction of disease, poverty, interpersonal conflict, and other catastrophic risks such as climate change. However, superintelligence, and advanced AI more generally, may also generate catastrophic vulnerabilities, including extreme inequality, global tyranny, instabilities that spark global (nuclear) conflict, catastrophically dangerous technologies, or, more generally, insufficiently controlled or aligned AI. Even if we successfully cross such safety and political pitfalls, tremendous governance questions confront us related to *what we want*, and what *we ought to want*, the answers to which will require us to know ourselves and our values much better than we do today.

Overview

AI Governance can be organized in several ways. This agenda divides the field into three complementary research clusters: the **technical landscape**, **politics**, and **ideal governance**. Each of these clusters characterizes a set of problems and approaches, within which the density of conversation is likely to be greater. However, most work in this space will need to engage the other clusters, drawing from and contributing high-level insights. This framework can perhaps be clarified by analogy to the problem of building a new city. The **technical landscape** examines the technical inputs and constraints to the problem, such as trends in the price and strength of steel. **Politics** considers the contending motivations of various actors (such as developers, residents, businesses), the possible mutually harmful dynamics that could arise and strategies for cooperating to overcome them. **Ideal Governance** involves understanding the ways that infrastructure, laws, and norms can be used to build the best

[17] Armstrong, Stuart, Nick Bostrom, and Carl Shulman. “Racing to the Precipice: A Model of Artificial Intelligence Development.” Technical Report. Future of Humanity Institute, 2013.

city, and proposing ideal master plans of these to facilitate convergence on a common good vision.

The first cluster, the **technical landscape**, seeks to understand the technical inputs, possibilities, and constraints, serving as a foundation for the other clusters of AI strategy. This includes **mapping** what could be the capabilities and properties of advanced and transformative AI systems, when particular capabilities are likely to emerge, and whether they are likely to emerge gradually in sequence or rapidly across-the-board. To the extent possible this cluster involves **modeling AI progress**: what the production function of AI progress is, given inputs such as compute, talent, data, and time. This cluster involves conceptualizing, measuring and projecting the relevant inputs, and **forecasting** progress to the extent possible, or else learning the extent to which forecasting is infeasible. We need to assess the viability, constraints, costs, and properties of scalably **safe AI systems**. To what extent will we need to invest resources and time late in the development process? What institutional arrangements best promote AI safety? To what extent will the characteristics of safe AI be apparent to and observable by outsiders, as would be necessary for (non-intrusive) external oversight and verification agreements?

The second cluster concerns **AI politics**, which focuses on the political dynamics between firms, governments, publics, researchers, and other actors, and how these will be shaped by and shape the technical landscape. How could AI transform **domestic and mass politics**? Will AI-enabled surveillance, persuasion, and robotics make totalitarian systems more capable and resilient? How will countries respond to the potentially massive increases in inequality and unemployment, and how will these responses support or hinder other global governance efforts? When and how will various actors become concerned and influential (what could be their "AI Sputnik" moments)? How could AI transform the **international political economy**? Will AI become regarded as the commanding heights of the modern economy, warranting massive state support and intervention? If so, what policies will this entail, which countries are best positioned to seize AI economic dominance, and how will this AI nationalism interact with global free trade institutions and commitments?

Potentially most importantly, how will AI interact with **international security**? What are the **near-term security challenges** (and opportunities) posed by AI? Could AI radically shift key strategic parameters, such as by enabling powerful new capabilities (in cyber, lethal autonomous weapons [LAWs], military intelligence, strategy, science), by shifting the offense-defense balance, by making crisis dynamics unstable, unpredictable, or more rapid? Could trends in AI facilitate new forms of international cooperation, such as by enabling strategic advisors, mediators, surveillance architectures, or by massively increasing the gains from cooperation and costs of non-cooperation? If general AI becomes regarded as a critical military (or economic) asset, could the state **control, close, and securitize** AI R&D, and if so how is it likely to proceed? What are the conditions that could spark and fuel an international **AI race**? How great are the dangers from such a race, how can those dangers be communicated and understood, and what factors could reduce or exacerbate the dangers? What routes exist for avoiding or escaping the race, such as norms, agreements, or institutions related to standards, verification, enforcement, or international control? How much does it matter to the world whether the leader has a large lead-margin, whether the leader is (based in) a particular country (e.g. US or China), or is governed in a particular way (e.g. transparently, by scientists)?

In steering away from dangerous rivalrous dynamics it will be helpful to have a clear sense of what we are steering towards, which bring us to the final research cluster: what are the **ideal governance** systems for global AI dynamics? What would we cooperate to build if we could? What potential global governance systems--including norms, policies, laws, processes, and institutions**--**can best ensure the beneficial development and use of advanced AI systems? To answer this we need to know what **values** humanity would want our governance system to pursue, or would want if we understood ourselves and the world better. More pragmatically, what are the specific interests of powerful stakeholders, and what **institutional mechanisms** exist to assure them of the desirability of the governance regime? Insights for long-term global governance are relevant to contemporary and medium-term AI governance, as we would like to embed today, when the stakes are relatively low, the principles and institutional mechanisms that will be crucial for the long-term. It will also facilitate cooperation today if we can assure powerful actors of a long-term plan that is compatible with their interests. A candidate policy proposal is that we want our AI leaders to constitutionally commit to the common good, to be sufficiently transparent, to demonstrate to stakeholders and others that they are pursuing the common good, to be exemplary in pursuing the common good and beneficial AI (which includes developing and employing best practices), and to be sufficiently accountable to prevent malign drift and hijacking.

In working in this space across the three research clusters, researchers should prioritize questions that seem most **important** (defined as most likely to yield beneficial insight), **tractable**, and **neglected**, and for which they have a **comparative advantage** and **interest**. Questions are more likely to be important if they are likely to identify or address crucial considerations, or if they directly inform urgent policy decisions. Questions are more likely to be tractable if the researcher can articulate a promising well-defined research strategy, the questions can be tackled in isolation, or they reduce to resolvable questions of fact. Questions are more likely to be neglected as they do not directly and exclusively contribute to an actor's profit or power, such as many long-term or global issues, and as they fall outside of the focus of traditional research communities. Researchers should upweight questions for which, comparatively, they have relevant expertise or capabilities, and in which they are especially interested. Ultimately, perhaps the simplest rule of thumb is to just begin with those questions or ideas that most grab you, and start furiously working.

Other Overviews

A rich overview to issues in the field is given in Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press, 2014. See especially chapters 4 ('The Kinetics of an Intelligence Explosion'), 5 ('Decisive Strategic Advantage'), 11 ('Multipolar scenarios'), and 14 ('The strategic picture').

The Future of Life Institute offers a set of resources on Global AI Policy here: https://futureoflife.org/ai-policy/.

Technical Landscape

Work on the technical landscape seeks to understand the technical inputs, possibilities, and constraints for AI, providing an essential foundation for our later study of AI politics, ideal governance, and policy. This includes mapping what the capabilities and properties of transformative AI systems could be, when they are likely to emerge, and whether they are likely to emerge in particular sequences or many-at-once. This research cluster benefits from expertise in AI, economic modeling, statistical analysis, technology forecasting and the history of technology, expert elicitation and aggregation, scenario planning, and neuroscience and evolution.

1. Mapping Technical Possibilities

This cluster investigates the more abstract, imaginative problem area of mapping technical possibilities, and especially potentially transformative capabilities. Are we likely to see a rapid broad (and local?) achievement of many transformative capabilities? What kinds of transformative capabilities could plausibly emerge, and in what order? What are their strategic properties, such as being offense- or defense- biased, or democracy or autocracy valenced?

1.1 Rapid and Broad Progress?

A first issue concerns how rapid and general advances will be in AI. Some believe that progress will, at some point, allow for sudden improvements in AI systems' capabilities across a broad range of tasks. If so, much of the following proposed work on sequences and kinds of AI would be unproductive, since most transformative capabilities would come online at the same explosive moment. For this reason, this agenda draws initial attention to this question.

Rapid general progress could come about from several mechanisms, enumerated with some redundancy:

- ❖ (1a) Many important tasks may require a common capability, the achievement of which would enable mastery across all of them. For example, deep learning unlocked seemingly disparate capabilities, spanning image recognition, language translation, speech recognition, game playing, and others. Perhaps a substantial advance in "efficient meta-learning" or transfer learning could catalyze advances in many areas.

- (1b) Clusters of novel powerful technological capabilities that are likely to be unlocked in close proximity to each other, perhaps because they facilitate each other or depend on solving some common problem.
- (2a) **Complements**: Scientific and technological advances often depend on having several crucial inputs, each of which acts as a strong complement to the others. For example, the development of powered flight seems to have required sufficient advances in the internal-combustion engine.[18] Complementarities could lead to jumps in capabilities in several ways.
    - (i) **Unjammed bottlenecks**: There could be rapid alleviation of a crucial bottleneck. For example, we have seen sudden jumps in capabilities from the provision of a single large training dataset for a particular task. Similarly, the generation of a crucial training set for a generally applicable task could lead to a broad front of progress.
    - (ii) **Overhang**: There could be a latent reservoir of a crucial complement, that becomes suddenly unlocked or accessible.[19] For example, it could come from: **hardware overhang** in which there is a large reservoir of compute available to be repurposed following an algorithmic breakthrough; from abundant **insecure compute** that can be seized by an expansionist entity; from **insight overhang**, if there are general powerful algorithmic improvements waiting to be uncovered; or from **data overhang**, such as the corpus of digitized science textbooks waiting to be read, and the internet more generally.
    - (iii) **Complementary clusters of capabilities**: Advances in one domain of AI could strongly complement progress in other domains, leading to a period of rapid progress in each of these domains. For example, natural language

[18] Crouch, Tom D, Walter James Boyne et al. "History of flight." *Encyclopædia Britannica*, 2016. https://www.britannica.com/technology/history-of-flight/The-generation-and-application-of-power-the-problem-of-propulsion.

[19] Unjammed bottlenecks and overhangs are closely related perspectives on complements, focusing either on the last necessary input or an already achieved necessary input. For example, consider the progress function f(X,Y, Z)=MIN(X,Y,Z). If at baseline X=0, Y=1, Z=1 then progress in X from 0 to 1 would represent an unjammed bottleneck. If at baseline X=0, Y=0, Z=1, then Z could be regarded as a form of overhang.

understanding could make it cost-effective to efficiently create massive datasets for a variety of purposes from the internet, creation of these datasets could improve machine understanding of how many task domains in the world relate to each other, which could improve transfer learning between those domains, which could further improve natural language understanding.

- ❖ (2b) Rapid progress in a crucial bottleneck/complement of AI research. For example, we have seen sudden jumps in capabilities from the provision of a single large training dataset for a particular task. Similarly, the generation of a crucial training set for a generally applicable task could lead to a broad front of progress.
- ❖ (3) Substantial AI advances on tasks crucial for future AI R&D, permitting highly[20] recursive self-improvement. This might lead to an endogenous growth positive feedback process, sometimes called an "intelligence explosion",[21] where each generation of AI accelerates the development of the subsequent generation.
- ❖ (4) Radical increases in investment in AI R&D.
- ❖ (5) A large ratio of R&D costs to execution costs, so that once a particular capability is achieved it could be massively deployed. For example, the learning process could be highly compute intensive (such as with genetic algorithms), but once trained that same compute could be used to run thousands of instances of the new algorithm.

The argument for rapid general progress emerging from recursive self-improvement, and associated strategy, has been articulated by Eliezer Yudkowsky,[22] Nick Bostrom,[23] and is being elaborated by MIRI. The argument against (spatially local) rapid general progress has been expressed by Robin Hanson,[24] Paul Christiano,[25] AI Impacts and Katja Grace,[26] and Ben

[20] Autoregressive parameter persistently above 1.

[21] Good, I.J. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers*, edited by Franz L. Alt and Moris Rubinoff, 6:31–88. New York: Academic Press, 1964.

[22] Yudkowsky, Eliezer. "Intelligence Explosion Microeconomics." Machine Intelligence Research Institute, 2013. https://intelligence.org/files/IEM.pdf.

[23] Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press, 2014. Chapter 4.

[24] Hanson, Robin. "I Still Don't Get Foom." *Overcoming Bias* (blog), July 24, 2014. http://www.overcomingbias.com/2014/07/30855.html.

[25] Christiano, Paul. "Takeoff Speeds." *The Sideways View* (blog), February 24, 2018. https://sideways-view.com/2018/02/24/takeoff-speeds/.

[26] Grace, Katja. "Likelihood of Discontinuous Progress around the Development of AGI." *AI Impacts*, February 23, 2018. https://aiimpacts.org/likelihood-of-discontinuous-progress-around-the-development-of-agi/.

Goertzel,[27] and is implicit to most mainstream perspectives.[28] The skeptical position draws support from the fact that AI progress and technological progress tends to be gradual, piecemeal, uneven, and spatially diffuse. If this remains true for AI then we should expect some transformative capabilities to come online far before others.

1.2 Kinds, Capabilities, and Properties of AI

AI could be transformative in many ways. We should systematically think through the kinds of AI that could be developed, and what their capabilities and properties might be.[29] For scenarios where progress is not rapid and broad, it will also be useful to articulate probable sequences in AI capabilities, or necessary achievements, prior to particular kinds of transformative AI (TAI).[30]

Some examples of potentially transformative capabilities include AI that is superhuman in, or otherwise transformative of, particular areas such as cybersecurity, autonomous weapons, surveillance, profiling, lie-detection, persuasion and manipulation, finance, strategy, engineering, manufacturing, and other areas of science and technology. Such AI, if arriving unbundled from other transformative capabilities, is often called "narrow AI". In addition to producing new capabilities, AI could be transformative through incremental effects, such as incremental changes in the costs or performance of existing capabilities, to the point that it transforms industries and world order.[31]

[27] Goertzel, Ben. "Superintelligence: Fears, Promises and Potentials: Reflections on Bostrom's Superintelligence, Yudkowsky's From AI to Zombies,and Weaver and Veitas's 'Open-Ended Intelligence.'" *Journal of Evolution & Technology* 24, no. 2 (November 2015): 55–87. http://www.kurzweilai.net/superintelligence-fears-promises-and-potentials.

[28] See also the reading list, compiled by Magnus Vinding, on arguments against the hard take-off' hypothesis: https://magnusvinding.com/2017/12/16/a-contra-ai-foom-reading-list/.

[29] There is some work on "kinds of intelligence" that may speak to this. For an informal introduction, see Shanahan, Murray. "Beyond Humans, What Other Kinds of Minds Might Be out There?" *Aeon*, October 19, 2016. https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there; Shanahan, Murray. "The Space of Possible Minds" EDGE, May 18, 2018. https://www.edge.org/conversation/murray_shanahan-the-space-of-possible-minds. See also the CFI project on 'Kinds of Intelligence', at http://lcfi.ac.uk/projects/kinds-of-intelligence/, and specifically José Hernández-Orallo, "The Measure of All Minds", 2017, Cambridge University Press, http://allminds.org/. Also see NIPS 2017 symposium: http://www.kindsofintelligence.org/.

[30] 'Transformative AI' ('TAI') is defined here to refer to advanced AI that could lead to radical changes in wealth, power, or world order. It is thus a lower threshold of "transformativeness" than has been used by OpenPhil, which is AI that "precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution." Cf. Karnofsky, Holden. "Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity." Open Philanthropy Project, 2016. http://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity.

[31] A discussion of the positive aspects of this is in Harford, Tim."What We Get Wrong about Technology." *FT Magazine*, July 17, 2017. https://www.ft.com/content/32c31874-610b-11e7-8814-0ac7eb84e5f1. Negative possibilities also exist. For example, even without any specific capabilities that are especially transformative or novel, AI and associated trajectories could displace sufficient workers to generate a political economic crisis on the scale of the Great Depression. "I suspect that if current trends continue, we may have a third of men between the ages of 25 and 54 not working by the end of this half century, because this is a

To date AI systems remain *narrow*, in the sense that a trained system is able to solve a particular problem well, but lacks the ability to generalize as broadly as a human can. Further, advances in AI capabilities are highly uneven, relative to the distribution of human capabilities, and this trend seems likely to persist: game playing algorithms are vastly superhuman at some games, and vastly subhuman at others.[32] AI systems today are sometimes analogized as "idiot savants": they vastly outperform humans at some tasks, but are incompetent at other "simple" adjacent tasks. AI systems are approaching or are now superhuman[33] at translating between languages, categorizing images, recognizing faces, and driving cars, but they still can't answer what seem like simple common-sense questions such as Winograd Schemas.

It may be the case that many kinds of TAI will arrive far before AI has achieved "common sense" or a child's ability to generalize lessons to a new task domain. Many thinkers, however, think the opposite is plausible. They reason that there is plausibly some faculty of general intelligence, some core cognitive module, some common factor to most kinds of "few-shot" learning (learning from only a few examples). This general intelligence, once achieved at even merely the level of a four-year-old human, would enable AI systems to be built that quickly learn in new domains, benefiting from and directing their superhuman memory, processing speed, sensor arrays, access to information and wealth of stored information, and library of specialized systems. This artificial general intelligence (AGI)--AI that can reason broadly across domains--could then rapidly catalyze progress across the task space; this is sometimes called "seed AGI".[34] The concept of AGI is more strategically relevant to the extent that (1) the

trend that shows no sign of decelerating. And that's before we have ... seen a single driver replaced [by self-driving vehicles] ..., not a trucker, not a taxicab driver, not a delivery person. ... And yet that is surely something that is en route." Quoted in Matthews, Christopher. "Summers: Automation is the middle class' worst enemy." *Axios*, June 4, 2017. https://www.axios.com/summers-automation-is-the-middle-class-worst-enemy-1513302420-754facf2-aaca-4788-9a41-38f87fb0dd99.html.

[32] Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, et al. "Human-Level Control through Deep Reinforcement Learning." *Nature* 518, no. 7540 (February 2015): 529–33. https://doi.org/10.1038/nature14236. See Figure 3, p. 531.

[33] For many definitions of the task, but not all.

[34] Yudkowsky provides a helpful overview of the concept of general intelligence here: Yudkowsky, Eliezer. "General Intelligence." *Arbital*, n.d. https://arbital.com/p/general_intelligence/. See also Goertzel, Ben. "Artificial general intelligence: concept, state of the art, and future prospects." *Journal of Artificial General Intelligence* 5.1 (2014): 1-48. Note that AGI, as defined in this document and by Yudkowsky and Goertzel, is conceptually distinct from broad human level capabilities. One could in principle have an AGI with sub-(adult)-human reasoning, or non-AGI systems with many superhuman capabilities. However, it does seem plausible that in practice AGI will be a necessary and sufficient condition for human-level capabilities in nearly all domains, given (1) the ability of the general intelligence to call on all the other existing superhuman assets of machine intelligence, and (2) the

concept maps onto a cluster of capabilities that come as a bundle (likely as a consequence of general reasoning), (2) AGI has transformative implications, such as igniting rapid general progress, (3) AGI arrives early in the sequence of transformative capabilities.[35]

We would like to know more about the probable strategic properties of novel capabilities and kinds of AI. For example, could they **enhance cooperation** by giving advice, by mediating or arbitrating disputes, by identifying gains-from-cooperation amongst strangers? Could AI and cheap surveillance enable robust monitoring of compliance to agreements, and cryptographic systems that protect participants from exposing their sensitive information? Could AI enable **overcoming commitment problems**[36] through binding costly commitments that are hard-coded into AI-adjudicated contracts? To what extent will AI enabled capabilities be **defense-biased** (vs offense-biased), defined here as costing relatively more to attack than to defend, for a given goal? Broadly speaking, defense-biased technology makes a multipolar world more stable.[37] To what extent will new technologies be **destruction-biased**, defined here as making it relatively easy to destroy value (but potentially hard to capture value)? Do new AI capabilities provide **first-mover advantages**, so that actors have (economic or military) incentives to develop and deploy them quickly?[38] An extreme form of power advantage, which may be more likely from first-mover advantages and offense bias, is **decisive strategic advantage**: an advantage sufficient to "achieve complete world domination".[39] The strategic character, and the perceived strategic character, of future technology will shape the international landscape, determining how secure or vulnerable are

vast array of problems seeming to require general intelligence--there is scarce data and extreme interdependencies with other domains--that are otherwise unlikely to be solved by narrow AI. Note that "human-level AI in X", "high human-level AI in X", and "superhuman AI in X" can be used to characterize narrow AI systems.

[35] For example, contra (1) it could be that general reasoning comes in different flavors, and AI becomes vastly superhuman at some forms while still remaining subhuman at others. Contra (3), AGI could plausibly be much harder to achieve than super-surveillance AI, super-hacking AI, even super-inventing AI (e.g. with circuit design). (2) seems plausible.

[36] Particularly between great powers, who otherwise lack a powerful legal structure within which to make commitments.

[37] For example, Yann LeCunn (NYU Ethics of AI Conference) stated that he believes narrow defensive AI systems will dominate general AI systems, because of specialization; his logic implies that narrow offensive AI systems should also dominate general AI systems, suggesting a world where general AI cannot flourish without massive narrow AI defenses. Eric Schmidt (2:52:59 during talk at Future of War Conference) conjectured that cyber-AI systems will dominate on the defense. (For discussion of near-term defense vs offense bias, see Brundage, Miles, Shahar Avin, et al. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." *ArXiv:1802.07228 [Cs]*, February 20, 2018. http://arxiv.org/abs/1802.07228.)

[38] There are several kinds of first mover advantage, such as from the first to attack, or the first to develop a capability. Both can be understood as a form of offense bias, though there are subtleties in definition related to what kinds of symmetry are presumed to be present.

[39] Bostrom. *Superintelligence*; p 96.

great powers under the status quo, and how able they are to cooperate to overcome commitment problems and the security dilemma.

These questions of the potential strategic properties of AI can also be framed in a more general way. To what extent will (particular kinds of) AI be, or have the option of being made to be, **transparent**, **stabilizing/destabilizing**, **centralizing/decentralizing** of power, **politically valenced** (towards authoritarianism or democracy), or **wisdom-enhancing** (advisor AIs)? How likely is it that we will get some of these (e.g. wisdom AI) before others (e.g. decisive cyber first strike AI)?

In researching the possible strategic capabilities of AI, we must also ask how far our estimates and models can be relied upon. Will developments be **predictable** and **foreseeable** in their character? To what extent will AI be **dual use**, making it hard to distinguish between the development, training, and deployment of dangerous/destabilizing/military systems and safe/stabilizing/economic systems? To what extent will developments be predictable in their **timelines and sequencing**, and what are our best forecasts (see section 2.3)? If we can estimate roughly when strategically relevant or transformative thresholds are likely to be reached, or in what order, then we can formulate a better map of the coming transformations.

1.3 Other Strategic Technology

Many other novel technologies could play a strategic or transformative role. It is worth studying them to the extent that they pose transformative possibilities before TAI, that they shape key parameters of AI strategy on the road to TAI, or that they represent technological opportunities that could be unlocked by a super R&D AI. These technologies include: atomically precise manufacturing, cheap and agile robotics, synthetic biology, genetic and cognitive enhancement, cyber-innovations and dependencies, quantum computing, ubiquitous and potent surveillance, lie detection, and military capabilities such as anti-missile defense, hypersonic missiles, energy weapons, ubiquitous subsea sensor networks, etc.

2. Assessing AI Progress

The previous section, Mapping Technical Possibilities, tried to creatively envision longer-run transformative possibilities and the character of AI. This section seeks to be more precise and quantitative about assessing existing and future progress in AI. Can we improve our measures of AI inputs, investments, and performance? Can we model AI progress: the relationship between measurable inputs and indicators, and future AI innovation? Supplementing model-based forecasts with expert assessment, to what extent can we forecast AI progress?

2.1 Measuring Inputs, Capabilities, and Performance

What are the key categories of **input** to AI R&D, and can we measure their existing distribution and rates of change? Plausibly the key inputs to AI progress are computing power (compute),[40] talent, data, insight, and money. Can we sensibly operationalize these, or find useful proxies for them? What are the most important AI **capabilities** that we should be tracking? Can we construct sensible, tractable, strategically relevant measures of **performance**, that either track or predict transformative capabilities? Prior and existing metrics of performance are summarized and tracked by the Electronic Frontier Foundation.[41]

This measurement exercise should be disaggregated at the level of the strategically relevant actor. Who are the main organizations and countries involved, and what is the distribution of and rates of change in their inputs, capabilities, and performance? Later in this document we will ask about the strategic properties of these organizations and countries, such as their institutional configuration (e.g. legal structure, leadership selection process), goals (political, economic, other), and access to other strategic assets. As a relatively poorly understood and potentially pivotal actor, current research is especially seeking to better understand China's inputs, capabilities, and performance.[42]

[40] Hwang, T. "Computational Power and the Social Impact of Artificial Intelligence." 2018, 1–44. http://dx.doi.org/10.2139/ssrn.3147971. Hilbert, M., and P. López. "The world's technological capacity to store, communicate, and compute information." *Science* 332, issue 6025 (April 1, 2011). http://doi.org/10.1126/science.1200970.

[41] At https://www.eff.org/ai/metrics.

[42] E.g. one relevant question is whether China can become a decisive world leader in AI without becoming more scientifically open. Wagner, Caroline S., and Koen Jonkers. "Open Countries Have Strong Science." *Nature News* 550, no. 7674 (October 5, 2017). https://doi.org/10.1038/550032a, p. 32. For an overview of China's AI landscape, see Ding, Jeffrey. "Deciphering China's

2.2 Modeling AI Progress

As well as mapping technical possibilities, we want to be able to model progress in AI development towards these possibilities.

A modeling strategy is to look for robust trends in a particular input, such as compute/$.[43] The canonical hardware trend is Moore's Law.[44] Kurzweil[45] and Nordhaus[46] observe an impressively consistent exponential trend in computing performance given cost, beginning before Moore's Law.[47] From this some argue that the trend will continue. *AI Impacts* finds the recent doubling time in FLOPS/$ to be about 3-5 years, slower than the 25 year trend of 1.2 years.[48] Amodei and Hernandez note that over the past five years there appears to be an exponential increase in the total compute used to train leading AI applications, with a 3.5 month doubling time.[49]

More complex approaches could try to build causal and/or predictive models of AI progress (on particular domains) as a function of inputs of compute, talent, data, investment, time, and indicators such as prior progress and achievements (modeling the "AI production function"). To what extent does performance scale with training time, data, compute, or other fungible assets?[50] What is the distribution of breakthroughs given inputs, and what is the existing and

AI Dream: The context, components, capabilities, and consequences of China's strategy to lead the world in AI." Future of Humanity Institute, March 2018. https://www.fhi.ox.ac.uk/wp-content/uploads/Deciphering_Chinas_AI-Dream.pdf.

[43] Such an approach works best when we can (1) credibly extrapolate the trend in the input (which may not be true if there is a change in underlying dynamics), and (2) can map the input to outcomes that matter (which may not be true for lots of reasons).

[44] For the original paper, see Moore, G.E. "Cramming More Components Onto Integrated Circuits." *Electronics* 38, no. 8 (April 19, 1965): 82–85. https://doi.org/10.1109/JPROC.1998.658762. Cf. also Schaller, R. R. "Moore's Law: Past, Present and Future." *IEEE Spectrum* 34, no. 6 (June 1997): 52–59. https://doi.org/10.1109/6.591665; "Trends in the cost of computing." *AI Impacts*, March 10, 2015. https://aiimpacts.org/trends-in-the-cost-of-computing/.

[45] Kurzweil, Ray. "The Law of Accelerating Returns." *Kurzweilai* (blog), March 7, 2001. http://www.kurzweilai.net/the-law-of-accelerating-returns.

[46] Nordhaus, William D. "Are we approaching an economic singularity? Information technology and the future of economic growth." Cowles Foundation Discussion Paper No. 2021, September 2015. Figure 1, https://cowles.yale.edu/sites/default/files/files/pub/d20/d2021.pdf.

[47] Kurzweil's data shows the trend beginning in 1900, Nordhaus's data shows the trend beginning in 1940.

[48] Grace, Katja. "Recent Trend in the Cost of Computing." *AI Impacts,* November 11, 2017. https://aiimpacts.org/recent-trend-in-the-cost-of-computing/. Doubling time is equal to time to a 10x increase divided by 3.3 (because $\log_2(10)=3.3$).

[49] Amodei, Dario, and Danny Hernandez. "AI and Compute." *OpenAI* (blog), May 16, 2018. https://blog.openai.com/ai-and-compute/.

[50] Silver, D., A. Huang, C.J. Maddison, A. Guez, and L. Sifre. "Mastering the game of Go with deep neural networks and tree search." *Nature* 529 (January 28, 2016). http://doi.org/10.1038/nature16961.

likely future distribution of inputs? How quickly can these assets be bought? How easy is it to enter or leapfrog?

Modeling these inputs may yield insights on rates of progress and the key factors which slow or expedite this. What do these models imply for likely bottlenecks in progress? Does it seem likely that we will experience hardware or insight overhang?[51] Put differently, how probable is it that a crucial input will suddenly increase, such as with algorithmic breakthroughs, implying a greater probability of rapid progress? More generally, from the perspective of developers, how smooth or sudden will progress be?

Articulating theoretically informed predictions about AI progress will help us to update our models of AI progress as evidence arrives. The status quo involves experts occasionally making ad-hoc predictions, being correct or mistaken by unquantified amounts, and then possibly updating informally. A more scientific approach would be one where explicit theories, or at least schools of thought, made many testable and comparable predictions, which could then be evaluated over time. For example, can we build a model that predicts time until super-human performance at a task, given prior performance and trends in inputs? Given such a model, we could refine it to assess the kinds of tasks and contexts where it is likely to make especially good or bad predictions. From this can we learn something about the size of the human range in intelligence space, for different kinds of tasks?[52] If our models are accurate then we would have a useful forecasting tool; if they are not then we will have hard evidence of our ignorance. We should build models predicting other strategically relevant parameters, such as the ratio of training compute costs to inference/execution compute costs, or more generally the ratio of R&D costs to execution costs (costs of running the system).[53]

[51] For example, a relatively simple algorithmic tweak generated massive improvements in Atari game playing. It is plausible that there are many other such easily implementable algorithmic tweaks that an AI could uncover and implement. Bellemare, Marc G., Will Dabney, and Rémi Munos. "A distributional perspective on reinforcement learning." *ArXiv:1707.06887 [Cs]*, July 21, 2017. https://arxiv.org/abs/1707.06887.

[52] Alexander, Scott. "Where The Falling Einstein Meets The Rising Mouse." *Slate Star Codex* (blog), August 3, 2017. http://slatestarcodex.com/2017/08/02/where-the-falling-einstein-meets-the-rising-mouse/.

[53] These parameters are relevant to the scale of deployment of a new system, to predicting the kinds of actors and initiatives likely to be innovating in various domains, and to other aspects of AI governance.

2.3 Forecasting AI Progress

Using the above measurements and models, and with expert judgment, to what extent can we forecast the development of AI (inputs and performance)? There are several desiderata for good forecasting targets. Given such forecasting efforts we could ask, how well calibrated and accurate are different groups of experts and models for different kinds of forecasting problems? How best can we elicit, adjust, and aggregate expert judgment? How different are the problems of near-term and long-term forecasting, and to what extent can we use lessons from or performance in near-term forecasting to improve long-term forecasts?

Near-term forecasting work is currently being done by Metaculus.[54] Many surveys have asked untrained and uncalibrated experts about near- and long-term forecasts. It can also be productive to evaluate previous forecasting efforts, to see how well calibrated they are, and if there are conditions that make them more or less accurate.[55]

3. AI Safety[56]

3.1 The Problem of AI Safety

AI Safety focuses on the technical challenge of building advanced AI systems that are safe and beneficial. Just as today a lot of engineering effort goes into ensuring the safety of deployed systems--making sure bridges don't fall down, car brakes don't fail, hospital procedures administer the correct medications to patients, nuclear power plants don't melt, and nuclear bombs don't unintentionally explode[57]--so it is plausible that substantial effort will be required to ensure the safety of advanced and powerful AI systems.

[54] At https://www.metaculus.com/questions/.

[55] Muehlhauser, Luke. "Retrospective Analysis of Long-Term Forecasts". https://osf.io/ms5qw/register/564d31db8c5e4a7c9694b2be.

[56] We use *AI Safety* to refer to the distinct, specialized field focusing on technical aspects of building beneficial and safe AI. AI Safety and AI Governance can be used as exclusive (and exhaustive) categories for the work needed to build beneficial AI. This agenda summarizes the aspects of AI Safety especially relevant to AI Governance.

[57] Though even this often requires considerable organizational efforts, and involves many close calls; cf. Sagan, Scott D. *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons*. Princeton: Princeton University Press, 1993; Schlosser, Eric. *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety*. Reprint edition. New York: Penguin Books, 2014.

Relatively simple AI systems are at risk of accidents and unanticipated behavior. Classifiers are known to be fragile and vulnerable to subtle adversarial attacks. One military report characterizes AI (specifically deep learning) as "weak on the 'ilities'",[58] which include reliability, maintainability, accountability, verifiability, debug-ability, fragility, attackability. AI can behave in a manner that is not foreseen or intended, as illustrated by Microsoft's failure to anticipate the risks of its "Tay" chatbot learning from Twitter users to make offensive statements. Complex systems, especially when fast-moving and tightly coupled, can lead to emergent behavior and 'normal accidents'.[59] The 2010 "flash crash" is illustrative, in which automated trading algorithms produced 20,000 "erroneous trades" and a sudden trillion dollar decline in US financial market value; this undesired behavior was stopped not by real-time human intervention but by automated safety mechanisms.[60]

The previous kinds of accidents arise because the AI is "too dumb". More advanced AI systems will overcome some of these risks, but gain a new kind of accident risk from being "too clever". In these cases a powerful optimization process finds "solutions" that the researchers did not intend, and that may be harmful.[61] Anecdotes abound about the surprising routes by which artificial life "finds a way", from a boat-racing AI that

[58] Potember, Richard. "Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD." JASON - The MITRE Corporation, 2017. https://fas.org/irp/agency/dod/jason/ai-dod.pdf, p. 2.

[59] Perrow, Charles. *Normal accidents: Living with high risk technologies*. Princeton, NJ: Princeton University Press, 2011; for discussions of normal accidents in AI systems, see: Maas, Matthijs. "Regulating for 'Normal AI Accidents'—Operational Lessons for the Responsible Governance of AI Deployment." New Orleans: Association for the Advancement of Artificial Intelligence, 2018. http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_118.pdf; and for normal accidents in the specific context of military weapon systems, see Danzig, Richard. "Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority." Center for a New American Security, June 2018. https://s3.amazonaws.com/files.cnas.org/documents/CNASReport-Technology-Roulette-DoSproof2v2.pdf?mtime=2018062807210; Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War*. New York: W. W. Norton & Company, 2018, pp. 150-155; Scharre, Paul. "Autonomous Weapons and Operational Risk." Ethical Autonomy Project. 20YY Future of Warfare Initiative. Center for a New American Security, 2016. https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf; Borrie, J. "Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies." UNIDIR Resources. UNIDIR, 2016. http://www.unidir.org/files/publications/pdfs/safety-unintentional-risk-and-accidents-en-668.pdf.

[60] U.S. Commodity Futures Trading Commission, and U.S. Securities & Exchange Commission. "Findings Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues." 2010. https://www.sec.gov/news/studies/2010/marketevents-report.pdf, p. 104. Linton, Oliver, and Soheil Mahmoodzadeh. "Implications of high-frequency trading for security markets." *Annual Review of Economics* 10 (2018). https://www.annualreviews.org/doi/pdf/10.1146/annurev-economics-063016-104407.

[61] Stuart Russell writes: "A system that is optimizing a function of n variables, where the objective depends on a subset of size k<n, will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable. This is essentially the old story of the genie in the lamp, or the sorcerer's apprentice, or King Midas: you get exactly what you ask for, not what you want. A highly capable decision maker – especially one connected through the Internet to all the world's information and billions of screens and most of our infrastructure – can have an irreversible impact on humanity." Russell, Stuart. "Of Myths And Moonshine." *Edge*, 2014. https://www.edge.org/conversation/the-myth-of-ai#26015.

reward-hacked by driving in circles,[62] to a genetic algorithm intended to evolve an oscillating circuit that hacked its hardware to function as a receiver for radio waves from nearby computers,[63] to a tic-tac-toe algorithm that learned to defeat its rivals using a memory bomb.[64]

Future advances in AI will pose additional risks (as well as offer additional opportunities for safety). AI systems will be deployed in ever more complex and consequential domains. They will be more intelligent at particular tasks, which could undermine the value of human oversight. They will begin to acquire models of their environment, and of the humans and institutions they interact with; they will gain understanding of human motivation, be able to observe and infer human affect, and become more capable of persuasion and deception. As systems scale to and beyond human-level (in particular dimensions), they may increasingly be able to intelligently out-maneuver human built control systems.

This problem is analogous to the problem of alignment in capitalism (of "avoiding market failures"): how to build a legal and regulatory environment so that the profit motive leads firms to produce social value. History is replete with examples of large negative externalities caused by firms perversely optimizing for profit, from fraudulent profits produced through 'creative' accounting (e.g. Enron), to policies that risk disaster or generate pollution (e.g. Deepwater Horizon oil spill), to firms that actively deceive their investors (e.g. Theranos) or regulators (e.g. Volkswagen emissions scandal). These scandals occur despite the existence of informed humans with "common sense" within the corporations, and the governance institutions being capable of comparable intelligence. Powerful AI systems may lack even that common sense, and could conceivably be much more intelligent than their governing institutions.

[62] Amodei, Dario, and Jack Clark. "Faulty Reward Functions in the Wild." *OpenAI* (blog), 2016. https://openai.com/blog/faulty-reward-functions/.

[63] Bird, Jon and Paul Layzell. "The Evolved Radio and its Implications for Modelling the Evolution of Novel Sensors." *Proceedings of the 2002 Congress on Evolutionary Computation*, CEC'02 (Cat. No.02TH8600), 2002. See https://people.duke.edu/~ng46/topics/evolved-radio.pdf.

[64] Many other examples can be found in Lehman, Joel et al. "The Surprising Creativity of Digital Evolution." *ArXiv:1803.03453 [Cs]*, March 29, 2018. https://arxiv.org/abs/1803.03453.

3.2 AI Safety as a Field of Inquiry

Much of the initial thinking about AI safety was focused on the challenge of making hypothesized human-level or superhuman AI safe. This line of inquiry led to a number of important insights.[65]

1. Orthogonality thesis: Intelligent systems could be used to pursue any value system.[66]
2. Instrumental convergence: systems will have instrumental reasons for acquiring power and resources, maintaining goal integrity, and increasing its capabilities and intelligence.
3. Empirical safety tests may not be sufficient. Human overseers may not have the capacity to recognize problems due to the system's complexity, but also its ability to intelligently model and game the oversight mechanism.
4. Formalizing human preferences is hard. When such a formal statement is fed as the goal into powerful and intelligent systems, they are prone to fail in extreme ways. This failure mode is a trope in Western literature, as per Midas' curse, the Sorcerer's Apprentice, the Monkey's Paw, and the maxim to be careful what one wishes for.[67]
5. There are many ways control or alignment schemes could catastrophically and irreversibly fail, and among the most dangerous are those we haven't thought of yet.

The above framing adopts the lens of *AI accident risks*: the risks of undesired outcomes arising from a particular, intentionally designed, AI system (often highly intelligent). There is another, relatively neglected framing, of *AI systemic risks*: the risks of undesired outcomes--some of which may be very traditional--that can emerge from a system of competing and cooperating agents and can be amplified by novel forms of AI. For example, AI could increase the risk of inadvertent nuclear war, not because of an *accident* or *misuse*, but

[65] Bostrom. *Superintelligence*; Yudkowsky, Eliezer. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Cirkovic, pp. 308–45. New York: Oxford University Press, 2008.

[66] Bostrom. *Superintelligence*, pp. 105-108.

[67] See Soares, Nate. "Ensuring Smarter-than-Human Intelligence has a Positive Outcome." Talks at Google series, November 20, 2016. https://www.youtube.com/watch?v=dY3zDvoLoao. Bostrom. *Superintelligence*, Chapter 9. Yudkowsky, Eliezer. "Difficulties of AGI Alignment." The Ethics of Artificial Intelligence Conference, NYU , 2016. https://livestream.com/nyu-tv/ethicsofAI/videos/138893593.

because of how AI could rapidly shift crucial strategic parameters, before we are able to build up compensating understandings, norms, and institutions.[68]

AI safety can thus be understood as the technical field working on building techniques to reduce the risks from advanced AI. This includes the ultimate goals of safety and alignment of superintelligent systems, the intermediate goals of reducing accident, misuse, and emergent risks from advanced systems, as well as near-term applications such as building self-driving car algorithms that are sufficiently safe, including being resilient to en-masse terrorist hacks.

As evidence of the importance of this field, when AI researchers were surveyed about the likely outcome of super-human AI, though the majority believe it is is very likely to be beneficial, the majority of respondents assign at least a 15% chance that superhuman AI would be "on balance bad" or worse, and at least a 5% chance it would be "extremely bad (e.g. human extinction)".[69] The goal of AI safety is to provide technical insights, tools, and solutions for reducing the risk of bad, and especially extremely bad, outcomes.

As AI systems are deployed in ever more safety-critical and consequential situations, AI researchers and developers will increasingly confront safety, ethical, and other challenges. Some solutions to these challenges will be one-off, local patches. For example, Google's solution to misclassifying images of black people as "gorillas" was to simply remove "gorilla" and similar primate categories from its service.[70] This kind of patch will not scale or generalize.

We would prefer to find solutions that are more foundational or generalizable, and thus more plausibly contribute to **scalably** safe and beneficial AI. Broadly, for particular systems we will want them to have various desirable properties, such as the following (drawing from Everitt et al's 2018 framework):

[68] Horowitz, Michael C., Paul Scharre, and Alex Velez-Green. "A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence." Working Paper, December 2017; Geist, Edward, and Andrew J Lohn. "How Might Artificial Intelligence Affect the Risk of Nuclear War?" RAND, 2018. https://www.rand.org/pubs/perspectives/PE296.html.
[69] Grace et al. (2017), "When Will AI Exceed Human Performance? Evidence from AI Experts."
[70] Simonite, Tom. "When It Comes to Gorillas, Google Photos Remains Blind." WIRED, January 11, 2018. https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/.

- **Reliability and Security**, so that the system behaves as intended in a wide range of situations, including under adversarial attack.[71]
- **Corrigibility**, so that the system is optimally open to being corrected by a human overseer if it is not perfectly specified/trained.[72] Candidate methods include by making the agent indifferent to or ignorant of interventions,[73] or uncertain about the reward function.[74]
- **Intelligibility**, interpretability, and transparency, such as through dimensionality reduction,[75] natural language communication, and techniques for visualizing or otherwise understanding what features parts of the learned algorithm are encoding.[76]
- **Value specification**, related to alignment,[77] formalizing current ethical principles,[78] inverse reinforcement learning and learning human preferences,[79] overcoming reward corruption,[80] and measuring and minimizing extreme side effects.[81]

[71] Adversarial examples: Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy (2014). "Explaining and Harnessing Adversarial Examples." *ArXiv: 1412.6572 [Stat],* March 20, 2015. https://arxiv.org/abs/1412.6572. Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. "Synthesizing Robust Adversarial Examples." *ArXiv:1707.07397 [Cs]*, July 24, 2017. http://arxiv.org/abs/1707.07397. Brown, Tom B., Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. "Adversarial Patch." *ArXiv:1712.09665 [Cs]*, December 27, 2017. http://arxiv.org/abs/1712.09665. Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images." *ArXiv:1412.1897 [Cs]*, December 5, 2014. http://arxiv.org/abs/1412.1897.

[72] Nate Soares, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. "Corrigibility." AAAI 2015 Ethics and Artificial Intelligence Workshop, 2015. https://intelligence.org/files/Corrigibility.pdf.

[73] Orseau, Laurent, and Stuart Armstrong. "Safely Interruptible Agents." October 28, 2016. https://intelligence.org/files/Interruptibility.pdf; Everitt, Tom, Daniel Filan, Mayank Daswani, and Marcus Hutter. "Self-Modification of Policy and Utility Function in Rational Agents." *ArXiv:1605.03142 [Cs]*, May 10, 2016. http://arxiv.org/abs/1605.03142; Armstrong, Stuart, and Xavier O'Rourke. "'Indifference' Methods for Managing Agent Rewards." *ArXiv:1712.06365 [Cs]*, December 18, 2017. http://arxiv.org/abs/1712.06365.

[74] Hadfield-Menell, Dylan, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. "Inverse Reward Design." *ArXiv:1711.02827 [Cs]*, November 7, 2017. http://arxiv.org/abs/1711.02827.

[75] Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, et al. "Human-Level Control through Deep Reinforcement Learning." *Nature* 518, no. 7540 (February 26, 2015): 529–33. https://doi.org/10.1038/nature14236.

[76] Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. "Feature Visualization." *Distill* 2, no. 11 (November 7, 2017): e7. https://doi.org/10.23915/distill.00007; Olah, Chris, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. "The Building Blocks of Interpretability." *Distill* 3, no. 3 (March 6, 2018): e10. https://doi.org/10.23915/distill.00010.

[77] Bostrom. *Superintelligence*.

[78] 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; Baum, Seth D. "Social Choice Ethics in Artificial Intelligence." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, October 2, 2017. https://papers.ssrn.com/abstract=3046725.

[79] Christiano, Paul, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. "Deep Reinforcement Learning from Human Preferences." *ArXiv:1706.03741 [Cs, Stat]*, June 12, 2017. http://arxiv.org/abs/1706.03741; Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. "Cooperative Inverse Reinforcement Learning." *ArXiv:1606.03137 [Cs]*, June 9, 2016. http://arxiv.org/abs/1606.03137; Evans, Owain, Andreas Stuhlmueller, and Noah D. Goodman. "Learning the Preferences of Ignorant, Inconsistent Agents." *ArXiv:1512.05832 [Cs]*, December 17, 2015. http://arxiv.org/abs/1512.05832; Choi, Jaedeug, and Kee-Eung Kim. "Inverse Reinforcement Learning in Partially Observable Environments." *Journal of Machine Learning Research* 12 (2011): 691–730. http://www.jmlr.org/papers/volume12/choi11a/choi11a.pdf ; Ng, Andrew Y., and Stuart J. Russell. "Algorithms for Inverse Reinforcement Learning." In *Proceedings of the Seventeenth International Conference on Machine Learning*, 663–70, 2000. https://ai.stanford.edu/~ang/papers/icml00-irl.pdf.

[80] Everitt, Tom, Victoria Krakovna, Laurent Orseau, Marcus Hutter, and Shane Legg. "Reinforcement Learning with a Corrupted Reward Channel." *ArXiv:1705.08417 [Cs]*, May 23, 2017. http://arxiv.org/abs/1705.08417.

- **Limiting capabilities**, such as through boxing, preferring 'oracle' AIs,[82] or building AI services rather than general AI agents.[83]
- **Performance and safety guarantees**, such as formal verification to identify upper bounds on the probability of unsafe behaviour or restrictions on exploration policies.

To some extent these approaches can be trialed and developed in concrete near-term settings.[84]

3.3 The Implications of AI Safety for AI Governance

For the purposes of AI governance it is important that we understand the strategic parameters relevant to building safe AI systems, including the viability, constraints, costs, and properties of scalably safe systems. What is the **safety production function**, which maps the impact of various inputs on safety? Plausible inputs are compute, money, talent, evaluation time, constraints on the actuators, speed, generality, or capability of the deployed system, and norms and institutions conducive to risk reporting. To what extent do we need to spend time or resources at various stages of development (such as early or late) in order to achieve safety? If the safety-performance trade-offs are modest, and political or economic returns to absolute and relative performance are relatively inelastic (marginal improvements in performance are not that important), then achieving safe AI systems is more likely to be manageable; the world will not have to resort to radical institutional innovation or other extreme steps to achieve beneficial AI. If, however, the safety-performance trade-off is steep, or political or economic returns are highly elastic in absolute or especially relative performance, then the governance problem will be much harder to solve, and may require more extreme solutions.

[81] Krakovna, Victoria, Laurent Orseau, Miljan Martic, and Shane Legg. "Measuring and Avoiding Side Effects Using Relative Reachability." *ArXiv:1806.01186 [Cs, Stat]*, June 4, 2018. http://arxiv.org/abs/1806.01186.

[82] Bostrom. *Superintelligence*, p. 145.

[83] Drexler, Eric. " Development-oriented models of superintelligence: Overview and topical documents." Work in progress. http://bit.ly/DrexlerAI.

[84] 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; Leike, Jan, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. "AI Safety Gridworlds." *ArXiv:1711.09883 [Cs]*, November 27, 2017. http://arxiv.org/abs/1711.09883.

There are a broad range of implicit views about how technically hard it will be to make safe advanced AI systems. They differ on the technical difficulty of safe advanced AI systems, as well as risks of catastrophe, and rationality of regulatory systems. We might characterize them as follows:

- **Easy**: We can, with high reliability, prevent catastrophic risks with modest effort, say 1-10% of the costs of developing the system.
- **Medium:** Reliably building safe powerful systems, whether it be nuclear power plants or advanced AI systems, is challenging. Doing so costs perhaps 10% to 100% the cost of the system (measured in the most appropriate metric, such as money, time, etc.).
    - **But incentives are aligned**. Economic incentives are aligned so that companies or organizations will have correct incentives to build sufficiently safe systems. Companies don't want to build bridges that fall down, or nuclear power plants that experience a meltdown.
    - **But incentives will be aligned**. Economic incentives are not perfectly aligned today, as we have seen with various scandals (oil spills, emissions fraud, financial fraud), but they will be after a few accidents lead to consumer pressure, litigation, or regulatory or other responses.[85]
    - **But we will muddle through**. Incentives are not aligned, and will never be fully. However, we will probably muddle through (get the risks small enough), as humanity has done with nuclear weapons and nuclear energy.
    - **And other factors will strongly work against safety**. Strong profit and power incentives, misperception, heterogenous theories of safety, overconfidence and rationalization, and other pathologies conspire to deprive us of the necessary patience and humility to get it right. This view is most likely if there will not be evidence (such as recoverable accidents) from reckless development, and if the safety function is steep over medium level of inputs ("This would not be a hard problem if we had two years to work on it, once we have the system. It will be almost impossible if we don't.").

[85] This assumes that recoverable accidents occur with sufficient probability before non-recoverable accidents.

- **Hard or Near Impossible**: Building a safe superintelligence is like building a rocket and spacecraft for a moon-landing, without ever having done a test launch.[86] It costs greater than, or much greater than, 100% of development costs.
- **We don't know**.

Will we be able to correctly diagnose the character of the steps required for sufficient safety, in development and deployment? Will we be able to agree on a common safety policy? Will we be able to verify compliance with that policy? For example, would it be possible to separate a machine's objectives from its capabilities, as doing so could make it easier for non-experts to politically evaluate a system and could enable verification schemes that leak fewer technical secrets (related to capabilities)?

Greater insight into the character of the safety problem will shed insight into a number of parameters relevant to solving the governance problem. Some governance arrangements that could depend on the character of the safety problem include:

- Providing incentives and protections for whistleblowers
- Representation of AI scientists in decision making
- Technical verification of some properties of systems
- Explicit negotiations over the goals of the system

AI Safety work is being done at a number of organizations, including DeepMind, OpenAI, Google Brain, the Center for Human Compatible AI and UC Berkeley, the Machine Intelligence Research Institute, the Future of Humanity Institute, and elsewhere.

[86] Yudkowsky, Eliezer. "So Far: Unfriendly AI Edition." EconLog | Library of Economics and Liberty, 2016. http://econlog.econlib.org/archives/2016/03/so_far_unfriend.html.

AI Politics

AI will transform the nature of wealth and power. The interests and capabilities of powerful actors will be buffeted, and new powerful actors may emerge. These actors will compete and cooperate to advance their interests. Advanced AI is likely to massively increase the potential gains from cooperation, and potential losses from non-cooperation; we thus want political dynamics to be such as to be most likely to identify opportunities for mutual benefit and to identify far in advance joint risks that could be avoided by prudent policy.

Political dynamics could also pose catastrophic risks short of human-level AI if, for example, they lead to great power war or promote oppressive totalitarianism. Political dynamics will affect what considerations will be most influential in the development of (transformative) AI: corporate profit, reflexive public opinion, researchers' ethics and values, national wealth, national security, sticky international arrangements, or enlightened human interest. It is thus critical that we seek to understand, and if possible, beneficially guide, political dynamics.

AI Politics looks at how the changing technical landscape could transform **domestic and mass politics**, **international political economy**, and **international security**, and in turn how policies by powerful actors could shape the development of AI. Work in this cluster benefits from expertise in domestic politics, international relations, and national security, among other areas. It will involve a range of approaches, including theory (mathematical and informal), contemporary case studies, historical case studies, close contact with and study of the relevant actors, quantitative measurement and statistical analysis, and scenario planning.

4. Domestic and Mass Politics

AI has the potential to shape, and be shaped by, domestic and mass politics. As AI and related technologies alter the distribution of domestic power, **forms of government** will alter. This could mean a shift in power towards actors with the capital and authority to deploy powerful AI systems, such as elites, corporations, and governments. On the other hand, AI could be used to enhance democracy, for example through aligned personal digital assistants, surveillance architectures that increase the accountability of authorities, or decentralized (crypto-economic) coordination technologies. The impact of exacerbated **inequality and job displacement** on trends such as liberalism, democracy, and globalization could be substantial. What systems are possible for mitigating inequality and job displacement, and will they be sufficient? More generally, **public opinion** can be a powerful force when it is mobilized. Can we foresee the contours of how public opinion is likely to be activated and expressed? Will certain groups--cultures, religions, economic classes, demographic categories--have distinct perspectives on AI politics? This set of questions is generally less relevant to short timelines (e.g. AGI comes within 10 years).

4.1 Forms of Government

Domestic political structures, such as whether a government is **accountable** through elections and is **transparent** through public legislative debates and an informed free press, arise as a complex function of many factors, some of which will plausibly be altered by advanced AI. Some factors that seem especially important to determining the character of government, and in particular the extent to which it is liberal and democratic, are: 1) the (unequal) distribution of control over economic assets and wealth; (2) surveillance technologies and architectures; (3) repression technologies; (4) persuasion technologies; (5) personal advisor technologies; (6) collective action technologies.

Research on forms of government will examine plausible AI-driven trends in these and other factors, and evaluate possible strategies for mitigating adverse trends. This matters for extreme stakes because (i) trends in domestic governance speak to long-term trends in regime-type (e.g. democracy); (ii) it could influence the character of key actors in AI strategy, such as the character of the Chinese and US governments; (iii) it will inform the kinds of global institutions that will be feasible and their properties.

4.1.1 Inequality

There is an extensive and active literature on inequality and government. This literature should be reviewed, and lessons applied to our understanding about future forms of government, given trends in inequality (see section 4.2).

4.1.2 Surveillance

To what extent will AI and sensor technology enable cheap, extensive, effective surveillance? It is plausible that sufficient information about an individual's behavior, intent, and psychology--and of an individual's social network--will soon be generated through passive interactions with digital systems, such as search queries, emails, systems for affect and lie detection, spatial tracking through MAC addresses, face recognition, or other kinds of individual recognition. If so, a first order effect seems to be to shift power towards those entities who are able to use such information, plausibly to reinforce government authority, and thus authoritarian systems. However, super-surveillance could also prove beneficial, such as for enabling AI verification agreements and for enabling "stabilization" (the prevention of world-destroying technology). In addition, it may be possible to design AI-enabled surveillance in ways that actually reinforce other values and institutions, such as liberalism and democracy. For example, it may be possible to attenuate the typical tradeoff between security and privacy through cryptographically enabled privacy-preserving surveillance.[87]

4.1.3 Repression and Persuasion

Profiling of individuals, mapping of social networks, ubiquitous surveillance and lie detection, scalable and effective persuasion, and cost-effective autonomous weapons could all radically shift power to states. These trends may enable a state that is willing to do so to monitor and disrupt groups working against it. Autonomous weapons could permit a dictator to repress without requiring the consent of military officers or personnel, which have historically been one check on leaders. These trends should be mapped, understood, their potential consequences studied, and governance safeguards proposed.

[87] One creative idea is to use secure multiparty computation or homomorphic encryption when analyzing data for evidence of criminal activity. These cryptographic technologies make it possible to perform such analysis without having access to the underlying data in an unencrypted form. See Trask, Andrew. "Safe Crime Detection: Homomorphic Encryption and Deep Learning for More Effective, Less Intrusive Digital Surveillance." *iamtrask*, June 5, 2017. https://iamtrask.github.io/2017/06/05/homomorphic-surveillance/; Garfinkel, Ben. "The Future of Surveillance Doesn't Need to Be Dystopian." Talk at Effective Altruism Global, June 9, 2018; and Yin, Awa Sun. "Introducing Open Mined: Decentralised AI." *Becoming Human: Artificial Intelligence Magazine,* August 3, 2017. https://becominghuman.ai/introducing-open-mined-decentralised-ai-18017f634a3f.
Ben Garfinkel provides an excellent review of the state-of-the-art in cryptographic systems and their implications for political, economic, and social institutions in Garfinkel, Ben. "Recent Developments in Cryptography and Potential Long-Term Consequences." Future of Humanity Institute, 2018.

4.1.4 Advisors and Collective Action

AI could also conceivably empower citizens, relative to the state or elites. Personal advisor AIs could allow individual citizens to engage politically in a more informed manner, and at lower cost to themselves; however, this level of AI capability seems like it might be close to AGI (likely to occur relatively late in the sequence of transformative developments). AI systems could facilitate collective action, such as if it becomes possible to assemble and mobilize a novel political coalition, and new political cleavages, through scraping of social media for expressions of support for neglected political positions. Individuals could express more complex political strategies, and more efficiently coordinate. For example, an American citizen (in a plurality voting system that strongly rewards coordination) might want to state that they would vote for a third party candidate if 40% of the rest of the electorate also agrees to do so.

Other kinds of narrow AI advisors could transform domestic politics. Efficient AI translation could facilitate cross-language communication and coordination. AI political filters could exacerbate political sorting (filter bubbles), or could elevate political discourse by helping users to avoid low-credibility news and more easily identify counter-arguments. Video and audio affect and sincerity/lie detection, if effective and trusted, could incentivize greater sincerity (or self-delusion).[88]

4.2 Inequality, Job Displacement, Redistribution

AI seems very likely to increase inequality between people (and probably also between countries: see section 5).[89] The digitization of products, because of low marginal costs, increases winner-take-all dynamics. AI dramatically increases the range of products that can be digitized. AI will also displace middle-class jobs, and near-term trends are such that the replacement jobs are lower paying.[90] Labor share of national incomes is decreasing; AI is

[88] Thanks to Carl Shulman for the above.

[89] Korinek, Anton, and Joseph E. Stiglitz. "Artificial intelligence and its implications for income distribution and unemployment." No. w24174. National Bureau of Economic Research, 2017. https://www8.gsb.columbia.edu/faculty/jstiglitz/sites/jstiglitz/files/w24174.pdf.

[90] For a review of forecasts of AI displacement of human jobs, see Winick, Erin. "Business Impact Every study we could find on what automation will do to jobs, in one chart." *MIT Technology Review*, January 25, 2018. https://www.technologyreview.com/s/610005/every-study-we-could-find-on-what-automation-will-do-to-jobs-in-one-chart/. The three more prominent forecasts are from: Nedelkoska, Ljubica and Glenda Quintini."Automation, Skills Use and Training", OECD Social, Employment and Migration Working Papers, No. 202, OECD Publishing, Paris, 2018. https://www.oecd-ilibrary.org/docserver/2e2f4eea-en.pdf?expires=1527369566&id=id&accname=guest&checksum=F85DCC6

likely to exacerbate this.[91] Ultimately, with human-level AI, the labor share of income should become ever smaller. Given that capital is more unequally distributed than labor value, an increase in capital share of income will increase inequality.

AI seems to be generating (or is at least associated with) new natural global monopolies or superstar firms: there's effectively only one search engine (Google), one social network service (Facebook), and one online marketplace (Amazon). The growth of superstar firms plausibly drives the declining labor share of income.[92] These AI (quasi-)monopolies and associated inequality are likely to increasingly become the target of redistributive demands. Another risk to examine, for "slow scenarios" (scenarios in which other forms of transformative AI do not come for many decades) is of an international welfare race-to-the-bottom, as countries race with each other to prioritize the needs of capital and to minimize their tax burden. Research in this area should measure, understand, and project trends in employment displacement and inequality. What will be the implications for the policy demands of the public, and the legitimacy of different governance models? What are potential governance solutions?[93]

D03FB399E86A59357199FABB1; Frey, Carl Benedikt, and Michael A. Osborne. "The Future of Employment: How Susceptible Are Jobs to Computerisation?" *Technological Forecasting and Social Change* 114 (2017): 254-280. https://www.sciencedirect.com/science/article/pii/S0040162516302244; Mankiya, J., Susan Lund, Michael Chui, Jacques Bughin, Jonathan Woetzel, Parul Batra, Ryan Ko, and Saurabh Sanghvi. "Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages." McKinsey & Company, December 2017. https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Future%20of%20Organizations/What%20the%20future%20of%20work%20will%20mean%20for%20jobs%20skills%20and%20wages/MGI-Jobs-Lost-Jobs-Gained-Report-December-6-2017.ashx. On efforts to theorize and forecast future labor displacement, see: Brynjolfsson, Erik, and Tom Mitchell. "What can machine learning do? Workforce implications." *Science* 358, no. 6370 (2017): 1530-1534. http://science.sciencemag.org/content/358/6370/1530.

[91] Dorn, David, Lawrence F. Katz, Christina Patterson, and John Van Reenen. "Concentrating on the Fall of the Labor Share." *American Economic Review* 107, no. 5 (2017): 180-85. Autor, David, and Anna Salomons. "Is automation labor-displacing? Productivity growth, employment, and the labor share." Brookings Papers on Economic Activity, 2018. https://www.brookings.edu/wp-content/uploads/2018/03/1_autorsalomons.pdf.

[92] Dorn et al, 2017.

[93] Automation's threat to the labor market may already be affecting politics. An analysis of survey data from 17 European countries between 2002 and 2012 finds that respondents whose jobs were more automotable expressed greater support for redistribution (Thewissen, Stefan and David Rueda. "Automation and the Welfare State: Technological Change as a Determinant of Redistribution Preferences." *Comparative Political Studies,* 2017. https://doi.org/10.1177/0010414017740600.) A similar study across 15 European democracies shows that those more exposed to automation shocks indicated greater support for nationalist and radical-right parties (Anelli, Massimo, Italo Colantone and Piero Stanig. "We Were The Robots: Automation in Manufacturing and
Voting Behavior in Western Europe." Working paper, 2018. http://www.italocolantone.com/research.html.) In the U.S., exposure to automation was positively correlated with support for Donald Trump in the 2016 Presidential Election at the electoral district level (Frey, Carl Benedikt, Chinchih Chen, and Thor Berger. "Political Machinery: Did Robots Swing the 2016 U.S. Presidential Election?" Oxford Martin School, July 2018. https://www.oxfordmartin.ox.ac.uk/publications/view/2576.)

4.3 Public Opinion and Regulation

Historically public opinion has been a powerful force in technology policy (e.g. bans on GMOs or nuclear energy) and international politics (e.g. Sputnik). Further, as these examples illustrate, public opinion is not simply a reflection of elite interest. In the case of Sputnik, the US intelligence community was well aware of Soviet progress, and the Eisenhower administration did not want to engage in a space race and tried to persuade the American public that Sputnik was not a significant development.[94] And yet, within months Sputnik had triggered a reorientation of US technology policy, including the legislative formation of ARPA (today DARPA). It could thus be helpful to study public opinion and anticipate movements in public opinion, as can be informed by scenario based surveys, and studying particular groups who have been exposed to instances of phenomena (such as employment shocks, or new forms of surveillance) that could later affect larger populations. What kinds of public reactions could arise, leading to overly reactive policy and regulation? Could regulating AI (or taxing AI companies) become a new target of political campaigns, as already seems to be happening in Europe? This area of research will also help policymakers know how best to engage public opinion when an event occurs (and in general). It will also help scholars to communicate the results and value of their work.

5. International Political Economy

The next set of questions in the AI Politics research cluster examines the international political dynamics relating to the production and distribution of wealth. Economic success with AI and information technology seems to exhibit substantial returns to scale (e.g. Google[95]) and agglomeration economies (e.g. Silicon Valley).[96] If these trends persist it could lead to an (even more extreme) international AI oligopoly, where a few firms capture most of the value from providing AI services. Are there any relevant aspects to the competitive dynamics

[94] Ryan, Amy and Gary Keeley. “Sputnik and US Intelligence: The Warning Record.” *Studies in Intelligence* 61:3 (2017). https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol-61-no-3/pdfs/sputnik-the-warning-record.pdf.

[95] Due in part to the “virtuous cycle” between AI capabilities which attracts customers, which increases one’s data, which improves one’s AI capabilities, and the high fixed costs of developing AI services and low marginal cost of providing them.

[96] On industry concentration in AI, see the following and references: Bessen, James E. “Information Technology and Industry Concentration.” Boston Univ. School of Law, Law and Economics Research Paper No. 17-41, December 1, 2017. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3044730.

between companies; for example, to what extent are AI innovations being patented, are patentable, or are held as trade secrets?

Countries lacking AI industries currently worry that they are being shut out of the most rewarding part of the global value chain.[97] Some in the US and Europe currently worry that China is coercively/unfairly leveraging its market power to strategically extract technical competence from Western firms, and this was arguably the motivation for the Trump trade war.[98] These concerns could lead to **AI mercantilism** or **AI nationalism**,[99] following from strategic-trade theory, where countries devote substantial resources to retaining and developing AI capabilities, and to supporting their AI national champions. To what extent are countries (e.g. Canada) able to internalize the returns on their AI investments, or does talent inevitably gravitate towards and benefit the existing leaders in AI (e.g. Silicon Valley)?[100] What lessons emerge from examining the partial analogies of other general purpose technologies and economy wide transformations such as computerization, electrification, and industrialization?

Countries and companies are searching for other ways to economically benefit from the AI transformed economy. They are searching for rewarding nodes in the value chain in which they can specialize. Countries are examining policy levers to capture more of the rents from AI oligopolies, and aspire to build up their own AI champions (such as the EU rulings against Apple and Google, and China's exclusion of Google and Facebook). How substantial of an advantage does China have, as compared with other advanced developed (mostly liberal democratic) countries, in its ability to channel its large economy, collect and share citizen data, and exclude competitors? What steps could and would the U.S. take to reinforce its lead? What are the possibilities and likely dynamics of an international economic AI race? Is it

[97] On "high development theory" as applied to data flows and assets, see: Weber, Steven. "Data, development, and growth." *Business and Politics* (2017): 1-27. https://www.cambridge.org/core/journals/business-and-politics/article/data-development-and-growth/DC04765FB73157C8AB76AB1742ECD38A.

[98] Barboza, David. "How This U.S. Tech Giant Is Backing China's Tech Ambitions." New York Times, August 4, 2017. https://www.nytimes.com/2017/08/04/technology/qualcomm-china-trump-tech-trade.html.

[99] Hogarth, Ian. "AI Nationalism." *Ian Hogarth* (blog), June 13, 2018. https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism.

[100] Recent opening of two DeepMind satellites in Canada, and the Pan-Canadian AI Strategy suggest that other informed actors believe non-central locations are worth investing in.

plausible that countries would support domestic or allied consortia of AI companies, so as to better compete in industries that appear to be naturally oligopolistic?

Technological displacement will impact countries differentially, and countries will adopt different policy responses. What will those be? If redistributing wealth and retraining becomes a burden on profitable companies, could there be AI capital flight and an international race "to the bottom" of providing a minimal tax burden? If so, could the international community negotiate (and enforce) a global tax system to escape this perverse equilibrium? Or are AI assets and markets sufficiently tax inelastic (e.g. territorially rooted) as to prevent such a race-to-the-bottom?

Research on international political economy is most relevant for scenarios where AI does not (yet) provide a strategic military benefit, as once it does the logic of international security will likely dominate, or at least heavily shape, economic considerations. However, many IPE related insights equally apply to the international security domain, so there is value in studying these common problems framed in terms of IPE.

6. International Security

AI and related technologies are likely to have important implications for national and international security. It is also plausible that AI could have strategic and transformative military consequences in the near and medium-term, and that the national security perspective could become dominant. First, studying the **near-term security challenges** is helpful for understanding the context out of which longer-term challenges will emerge, and enable us to seed long-term beneficial precedents. Longer-term, if general AI becomes regarded as a critical military (or economic) asset, it is possible that the state will seek to **control**, **close**, **and securitize** AI R&D. Further, the strategic and military benefits of AI may fuel international **race dynamics**. We need to understand what such dynamics might look like, and how such a race can be **avoided** or **ended**.

6.1 Near-term Security Challenges

In the coming years AI will pose a host of novel security challenges. These include international and domestic uses of autonomous weapons, and AI-enabled cyber-operations, malware, and political influence campaigns ("active measures"). Many of these challenges look like "lite" versions of potential transformative challenges, and the solutions to these challenges may serve as a foundation for solutions to transformative challenges.[101] To the extent the near-term and transformative challenges, or their solutions, are similar, it will be useful for us to be aware of and engage with them. For a recommended syllabus of readings on AI and International Security, see:

- Zwetsloot, Remco. "Artificial Intelligence and International Security Syllabus." Future of Humanity Institute, 2018. (link).

Some specific references worth looking at include:

- Brundage, M., S. Avin, et al. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." Future of Humanity Institute, 2018. [PDF].
- Horowitz, Michael, Paul Scharre, Gregory C. Allen, Kara Frederick, Anthony Cho and Edoardo Saravalle. "Artificial Intelligence and International Security." Centre for a New American Security, 2018. (link).
- Scharre, P. *Army of None: Autonomous Weapons and the Future of War*. New York: W. W. Norton & Company, 2018.
- Horowitz, M. "Artificial Intelligence, International Competition, and the Balance of Power." *Texas National Security Review*, 2018. [link] [PDF].

6.2 Control, Closing, and Securitization

Basic AI R&D is currently conducted in the open: researchers have a strong interest to publish their accomplishments to achieve recognition, and there is a strong ethos of scientific openness. Some AI R&D is semi-closed, conducted in private for-profit spaces; however, this tends to not be general AI R&D, but instead applications of existing techniques. This could plausibly change, if AI becomes perceived as catastrophically dangerous, strategic military, or

[101] For example, we can analyze how transnational self-governance regimes of private companies have emerged and why these efforts have succeeded or failed. This is particularly relevant as several AI companies have already introduced self-governance measures as well. Fischer, Sophie-Charlotte. 2018. "Reading List - Industry Self-Regulation/Security Governance"..

even strategic economic. To the extent an AI race is likely and catastrophic risks are increasing in a close race, as opposed to one where the leader has a large lead, it would arguably be preferable for AI leaders to take steps to prevent their capabilities from diffusing to others.

What are the different models for completely or partially closing AI research or assets (like compute)? What are their pros and cons?[102] At what point would and should the state be involved? What are the legal and other tools that the state could employ (or are employing) to close and exert control over AI companies? With what probability, and under what circumstances, could AI research and development be *securitized*--i.e., treated as a matter of national security--at or before the point that transformative capabilities are developed? How might this happen and what would be the strategic implications? How are particular private companies likely to regard the involvement of their host government, and what policy options are available to them to navigate the process of state influence? How are researchers likely to be involved? Can we learn from the study of the attempted closing and control of other technologies?

6.3 Race Dynamics

Advanced AI could convey extreme power and wealth to its possessors. If so, and in particular if it is expected to convey strategic military or economic benefits, then it is plausible that an (international) race dynamic could emerge. The defining feature of a technology race is that there are large gains from relative advantage. In such a circumstance actors have strong incentives to trade-off against other values (like safety, transparency, accountability, democracy) and opportunities, in order to increase the probability of gaining advantage. In particular, a worry is that it may be close to a necessary and sufficient condition for AI safety and alignment that there be a high degree of caution prior to deploying advanced powerful systems; however, if actors are competing in a domain with large returns to first-movers or relative advantage, then they will be pressured to choose a sub-optimal level of caution.

[102] Bostrom, Nick. "Strategic Implications of Openness in AI Development." *Global Policy*, February 2017. http://onlinelibrary.wiley.com/doi/10.1111/1758-5899.12403/full or http://www.nickbostrom.com/papers/openness.pdf; Krakovna, Victoria. "Clopen AI: Openness in different aspects of AI Development." *Deep Safety* (blog), August 1, 2016. https://vkrakovna.wordpress.com/2016/08/01/clopen-ai-openness-in-different-aspects-of-ai-development/.

Research on race dynamics involves a large set of questions and approaches. We will need to integrate and develop models of technology/arms races.[103] What are the distinctive features of an AI race, as compared with other kinds of races? What robust predictions can we make about that subfamily of races? Under what conditions are those races most dangerous or destructive? Specifically, a plausible and important proposition is that races are more dangerous the smaller the margin of the leader; is this a robust conclusion?[104] How do openness, accessibility of the research frontier, first-mover advantages, insecure compute, and other factors affect race dynamics? Given models of AI innovation, how confident can a lead team be about the performance of its rivals, and that it will be able to sustain a known lead? Given models of AI safety (such as the performance-safety tradeoffs and the time-schedule for safety investments), what is the expected risk incurred by race dynamics?

There are also questions about the strategies for retaining a lead or catching up. Are there tools available to the leading team that will allow it to retain a lead? For example, could a team retain its lead by closing off its research? What difference does it make if the leading team is a state, or closely supported by a state?

The potential for coalitions within a race merits study. What are the possibilities for alliances between leading groups or states to help them retain their lead? In light of states' interests in strong AI systems, current international agreements, and historic relationships, what configurations of state coalitions are likely and under what circumstances?

Historical precedents and analogies can provide insight, such as consideration of the arms race for and with nuclear weapons, other arms races, and patent and economic technology races. What about analogies to other strategic general purpose technologies and more gradual technological transformations, like industrialization, electrification, and computerization? In what ways do each of these fail as analogies?

[103] For a model on risks from AI races, see Armstrong, Stuart, Nick Bostrom, and Carl Shulman. "Racing to the precipice: a model of artificial intelligence development." *AI & Society* 31, no. 2 (2016): 201-206.
The broader modeling literature relevant to races is large, including economic models of auctions, patent races, and market competition.

[104] For a model of a technology race which is most intense when the racers are far away, see Hörner, Johannes. "A Perpetual Race to Stay Ahead." *Review of Economic Studies* (2004) 71, 1065–1088. http://users.econ.umn.edu/~holmes/class/2007f8601/papers/horner.pdf.

Finally it would be valuable to theorize the likely stages of an AI race and their characteristics (tempo, danger, security consequences). Can we map current behavior onto this framework? What is the current distribution of capabilities, talent, and investment?[105] To what extent do existing policy makers and publics perceive or invoke a race logic?[106] What kinds of events could spark or escalate a race, such as “Sputnik moments” for publics[107] or an “Einstein-Szilard letter” for leaders?[108]

6.4 Avoiding or Ending the Race

Given the likely large risks from an AI race, it is imperative to **examine possible routes for avoiding races or ending one underway**. The political solutions to global public bads are, in increasing explicitness and institutionalization: norms, agreements (“soft law”), treaties, or institutions. These can be bilateral, multilateral, or global. Norms involve a rough mutual understanding about what (observable) actions are unacceptable and what sanctions will be imposed in response. Implicit norms have the advantage that they can arise without explicit consent, but the disadvantage that they tend to be crude, and are thus often inadequate and may even be misdirected.[109] A hardened form of international norms is customary law, though absent a recognized international judiciary this is not likely relevant for great-power cooperation.[110]

Diplomatic agreements and treaties involve greater specification of the details of compliance and enforcement; when well specified these can be more effective, but require greater levels

[105] This is also asked in the Technical Landscape.
[106] Cave, Stephen, and Seán S. Ó hÉigeartaigh. “An AI Race for Strategic Advantage: Rhetoric and Risks.” In *AAAI / ACM Conference on Artificial Intelligence, Ethics and Society*, 2018. http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_163.pdf.
[107] AlphaGo so far being the closest thing to a Sputnik moment, though that mostly for people in China, Japan, South Korea, and other cultures where Go is esteemed.
[108] Cf. Grace, Katja. “Leó Szilárd and the Danger of Nuclear Weapons: A Case Study in Risk Mitigation.” Technical Report. Berkeley, CA: Machine Intelligence Research Institute, October 2015. https://intelligence.org/files/SzilardNuclearWeapons.pdf .
[109] Joseph Nye advocates for cyber-norms. Nye, Joseph S. “A Normative Approach to Preventing Cyberwarfare.” Project Syndicate, March 13, 2017. https://www.project-syndicate.org/commentary/global-norms-to-prevent-cyberwarfare-by-joseph-s--nye-2017-03. However, the case of cyber weapons may also point to some technologies that are much more limited in their potential to be controlled through arms control agreements. For an introduction, cf. Borghard, Erica D., and Shawn W. Lonergan. “Why Are There No Cyber Arms Control Agreements?” Council on Foreign Relations, January 16, 2018. https://www.cfr.org/blog/why-are-there-no-cyber-arms-control-agreements.
[110] Cf. Williamson, Richard. “Hard Law, Soft Law, and Non-Law in Multilateral Arms Control: Some Compliance Hypotheses.” *Chicago Journal of International Law* 4, no. 1 (April 1, 2003). https://chicagounbound.uchicago.edu/cjil/vol4/iss1/7.

of cooperation to achieve. Institutions, such as the WTO, involve establishing a bureaucracy with the ability to clarify ambiguous cases, verify compliance, facilitate future negotiations, and sometimes the ability to enforce compliance. International cooperation often begins with norms, proceeds to (weak) bilateral or regional treaties, and consolidates with institutions.

Some conjectures about when international cooperation in transformative AI will be more likely are when: (1) the parties mutually perceive a strong interest in reaching a successful agreement (great risks from non-cooperation or gains from cooperation, low returns on unilateral steps); (2) when the parties otherwise have a trusting relationship; (3) when there is sufficient consensus about what an agreement should look like (what compliance consists of), which is more likely if the agreement is simple, appealing, and stable; (4) when compliance is easily, publicly, and rapidly verifiable; (5) when the risks from being defected on are low, such as if there is a long "breakout time", a low probability of a power transition because technology is defense dominant, and near-term future capabilities are predictably non-transformative; (6) the incentives to defect are otherwise low. Compared to other domains, AI appears in some ways less amenable to international cooperation--conditions (3), (4), (5), (6)--but in other ways could be more amenable, namely (1) if the parties come to perceive existential risks from unrestricted racing and tremendous benefits from cooperating, (2) because China and the West currently have a relatively cooperative relationship compared to other international arms races, and there may be creative technical possibilities for enhancing (4) and (5). We should actively pursue technical and governance research today to identify and craft potential agreements.

Third-Party Standards, Verification, Enforcement, and Control

One set of possibilities for avoiding an AI arms race is the use of third party standards, verification, enforcement, and control. What are the prospects for cooperation through third party institutions? The first model, almost certainly worth pursuing and feasible, is an international "safety" agency responsible for "establishing and administering safety standards."[111] This is crucial to achieve common knowledge about what counts as compliance. The second "WTO" or "IAEA" model builds on the first by also verifying and ruling on

[111] As per Article II of the IAEA Statute.

non-compliance, after which it authorizes states to impose sanctions for noncompliance. The third model is stronger still, endowing the institution with sufficient capabilities to enforce cooperation itself. The fourth, "Atomic Development Authority" model, involves the agency itself controlling the dangerous materials; this would involve building a global AI development regime sufficiently outside the control of the great powers, with a monopoly on this (militarily) strategic technology. Especially in the fourth case, but also for the weaker models, great care will need to go into their institutional design to assure powerful actors, and ensure competence and good motivation.

Such third party models entail a series of questions about how such institutions could be implemented. What are the prospects that great powers would give up sufficient power to a global inspection agency or governing body? What possible scenarios, agreements, tools, or actions could make that more plausible? What do we know about how to build government that is robust against sliding into totalitarianism and other malignant forms (see section 4.1)? What can we learn from similar historical episodes, such as the failure of the Acheson-Lilienthal Report and Baruch Plan, the success of arms control efforts that led towards the 1972 Anti-Ballistic Missile (ABM) Treaty,[112] and episodes of attempted state formation?

There may also be other ways to escape the race. Could one side form a winning or encompassing coalition? Could one or several racers engage in unilateral "stabilization" of the world, without risking catastrophe? The section AI Ideal Governance discusses the desirable properties of a candidate world hegemon.

[112] Adler, Emanuel. "The Emergence of Cooperation: National Epistemic Communities and the International Evolution of the Idea of Nuclear Arms Control." *International Organization* 46, no. 1 (1992): 101–45.

AI Ideal Governance

The Technical Landscape seeks to understand the technical possibilities and constraints of AI development. **AI Politics** seeks to understand how different actors will compete and cooperate to achieve their objectives related to powerful AI. **AI Ideal Governance** focuses on cooperative possibilities: if we could sufficiently cooperate, what might we cooperate to build? AI Ideal Governance examines potential global arrangements for governing what kinds of AI are developed and deployed, by whom, for what purposes, and with what constraints. In particular, this cluster seeks to identify ideal models of global governance. What are humanity's common **values**, and what arrangements could best satisfy our distinct goals? What organizational principles and **institutional mechanisms** exist to best promote those? These age-old questions need to be investigated with renewed vigor. We may soon need to implement our best answer. Advanced AI could also dramatically alter the relevant parameters of the question, rendering prior insights less relevant.

This research cluster focuses on idealized global governance, for several reasons. It is important to devote thought to articulating what we are trying to achieve (1) so that we are best able to steer events in desirable directions and (2) to facilitate cooperation by coordinating on an appealing shared **vision**. In so doing it will be important to develop effective means of communicating the bounty of potential benefits from cooperation, and the existential dangers from rivalrous behavior. We need to make sure we understand the distinct concerns and worldviews of people and elites from different backgrounds, so that our governance proposals are most likely to resonate globally and to be incentive-compatible with powerful stakeholders. We need to think pragmatically about institutional design--what should be the constitutional foundation; who is entitled to make what decisions, under what conditions, by what voting rules; what information is transmitted to whom--to ensure that any acceptable ideal vision is politically stable.

The findings from this research cluster will be crucial for advising the governance of AI firms and countries today and in the future. This is so for two reasons.

(1) The governance problems that we are facing today and that we will face in the future overlap extensively, with the primary differences being (i) the scope of interests to be represented, (ii) the potential need to compete in some broader military-economic domain, and (iii) the stakes. To illustrate the similarities, consider how the governance of an international AI coalition will ideally have some constitutional commitment to a common good, will have institutional mechanisms for assuring the principals (e.g. the publics and leaders of the included countries) that the regime is well-governed, and for credibly communicating a lack of threat to other parties. In fact, if we are able to craft a sufficiently appealing, realistic, self-enforcing, robust model of AI governance, this could serve as a beacon, to guide us out of a rivalrous equilibrium. The problem then reduces to one of sequencing: how do we move from the present to this commonly appealing future?
(2) We would ideally like to embed into our governance arrangements today, when the stakes are relatively low, the principles and mechanisms that we will need in the future. For example, given temporal discounting, diminishing marginal utility, and uncertainty about who will possess the wealth, it may be possible today to institutionalize collective commitments for redistributing wealth in the future.[113]

This research cluster is the least developed in this document, and within the community of people working on AI governance.

7. Values and Principles

AI ideal governance aspires to envision, blueprint, and advance ideal institutional solutions for humanity's AI governance challenges. What are the common values and principles around which different groups can coordinate? What do various stakeholders (publics, cultural groups, AI researchers, elites, governments, corporations) want from AI, in the near-term and long-term? What are the best ways of mediating between competing groups and between conflicting values? What do long-term trends--such as from demographics, secularization, globalization, liberalism, nationalism, inequality--imply about the values of these stakeholders over medium and long timelines?

[113] Bostrom, Nick, Allan Dafoe, and Carrick Flynn. "Public Policy and Superintelligent AI: A Vector Field Approach." Future of Humanity Institute, 2018. http://www.nickbostrom.com/papers/aipolicy.pdf

Given what we are still learning about ourselves and our values, is it possible to anticipate the direction that our values are moving in, or the direction they should move in? Given uncertainty about our common values and what should be our values, are there principles we can employ that will "learn with us" over time and prevent us from making large mistakes? Bostrom, Dafoe, and Flynn (2018) offer a set of policy desiderata that gain distinctive importance in a world of superintelligent AI, including: expeditious progress, AI safety, conditional stabilization, non-turbulence, universal benefit, magnanimity, continuity, first-principles thinking, wisdom, speed and decisiveness, and adaptability. There appear to be two crucial (meta-)principles in the present world, and they are in tension: (1) **security** (AI safety, conditional stabilization) and (2) **autonomy** (freedom, continuity, sovereignty).

This work requires scholars of ethics and morality, psychology, global public opinion, culture, and religion.

8. Institutions and Mechanisms

The previous section involves specifying the interests of the stakeholders that a governance system should meet, as well as the overall goals of the system. This section then seeks to develop institutions that can successfully achieve these interests and goals.

We want our governance institutions to be capable of providing security, ensuring safety from non-aligned AI, and otherwise stabilizing technological development to prevent new extreme risks. This may require centralized control over AI development, or extensive surveillance of AI projects with ready ability to shut them down. It's possible such safety could be achieved through a cooperating multipolar world, but it may require concentration of power and authority. What are the least infringing possible stabilization arrangements? What capabilities may AI enable that could help us with this, how probable are they, and what could be done to increase their probability?

We want our governance institutions to be resilient to drift and hijacking. Two poles of the risk space are totalitarian capture and tyranny of the majority. To prevent totalitarian capture and tyranny of the majority, to varying extents and in varying combinations, countries throughout the world have employed: regular, free, and fair elections; protected rights for political expression; rule of law and an independent judiciary; division of power; constraints on state power; constitutionally protected rights; federalism.

The problem of how to build institutions for governing a polity is a core part of the fields of political science and political economy. The more mathematical theoretical corner of this space is often called public choice, social choice, or (by economists) political economy. Political scientists in comparative politics and American politics extensively study the properties of different political systems. Scholars in political science and political theory study the design of constitutions. Given the centrality of this problem to these fields, and their existing expertise, substantial effort should be spent learning from them and recruiting them, rather than trying to reinvent good governance. Nevertheless, at the present time the application of these disciplines to the problems of AI governance remains neglected.

9. Positive Visions

While the above is directed to devising a feasible model of ideal long-run AI governance, it is unlikely to generate a simple solution (anytime soon). However, it could be extremely beneficial to have a simple, compelling, broadly appealing vision of the benefits from cooperation, to help motivate cooperation. We believe that both the potential benefits from safe development and the potential downsides from unsafe development are vast. Given that perspective, it is foolish to squabble over relative gains, if doing so reduces the chances of safe development. How can we simply, clearly, evocatively communicate that vision to others?[114]

[114] “Paretotopia”?

Appendix A: Forecasting Desiderata

(This relates to section 2.3).

1. We want our forecasting targets to be indicators for **relevant** achievements. This includes targets that serve as (leading) indicators for important *economic* capabilities, such as a capability that would pose a substantial employment threat to a large group of people. It includes indicators for important *security* capabilities, such as in potent AI cyber-offense, surveillance and imagery intelligence, or lie detection. It includes indicators for important *technical* achievements, such as those that are thought to be crucial steps on the path to more transformative capabilities (e.g. AGI), those that are central to many problem areas, or those that would otherwise substantially accelerate AI progress.[115]
2. We want them to be **accurate** indicators, as opposed to noisy indicators that are not highly correlated with the important events. Specifically, where E is the occurrence or near occurrence of some important event, and Y is whether the target has been reached, we want P(not Y|not E)~1, and P(Y | E) ~1. An indicator may fail to be informative if it can be "gamed" in that there are ways of achieving the indicator without the important event being near. It may be a noisy indicator if it depends on otherwise irrelevant factors, such as whether a target happens to take on symbolic importance as the focus of research.
3. We want them to be **well specified**: they are unambiguous and publicly observable, so that it will not be controversial to evaluate whether E has taken place. These could be either targets based on a commonly agreed objective metric such as an authoritative measure of performance, or a subjective target likely to involve agreement across judges. Judges will not ask later: "what did you mean"?[116]
4. We want them to be somewhat **near-term probable**: we should not be too confident in the near-term about whether they will occur. If they all have tiny probabilities

[115] The Good Judgment Project sometimes refers to an indicator that is relevant as one that is *diagnostic* of a bigger issue that we care about.

[116] Tetlock has called this the "Clairvoyance Test": if you asked a clairvoyant about your forecasting question, would they be able to answer you or would they require clarification on what you meant. See Tetlock, Philip E., and Dan Gardner. Superforecasting: The art and science of prediction. Random House, 2016, and https://www.edge.org/conversation/philip_tetlock-how-to-win-at-forecasting

(<1%) then we will not learn much after not seeing any of them resolve.[117] The closer the probability of a forecasting event and a set of predictions is to 50%, over a given time frame, the more we will learn about forecasting ability, and the world, over that time frame.

5. We ideally want them to be **epistemically temporally fractal**: we want them to be such that good forecasting performance on near-term forecasts is informative of good forecasting performance on long-term predictions. Near-term forecasting targets are more likely to have this property as they depend on causal processes that are likely to continue to be relevant over the long-term.
6. We want them to be **jointly maximally informative**. This means that we ideally want a set of targets that score well on the above criteria. A way in which this could not be so is if some targets are highly statistically dependent on others, such as if some are logically entailed by others. Another heuristic here is to aim for forecasting targets that exhaustively cover the different causal pathways to relevant achievements.

[117] Though we should learn a lot from seeing one such unexpected event occur. Thus, such a "long-shot" target would be a worthwhile forecasting target to a person who assigns intermediate subjective probability of it occurring, even if everyone else in the community is confident it will (not) occur.