AI Governance: Opportunity and Theory of Impact

Allan Dafoe

September, 2020

AI governance concerns how humanity can best navigate the transition to a world with advanced AI systems[1]. It relates to how decisions are made about AI[2], and what institutions and arrangements would help those decisions to be made well.

I believe advances in AI are likely to be among the most impactful global developments in the coming decades, and that AI governance will become among the most important global issue areas. AI governance is a new field and is relatively neglected. I'll explain here how I think about this as a cause area and my perspective on how best to pursue positive impact in this space. The value of investing in this field can be appreciated whether one is primarily concerned with contemporary policy challenges or long-term risks and opportunities ("longtermism"); this piece is primarily aimed at a longtermist perspective. Differing from some other longtermist work on AI, I emphasize the importance of also preparing for more conventional scenarios of AI development.

Contemporary Policy Challenges

AI systems are increasingly being deployed in important domains: for many kinds of surveillance; by authoritarian governments to shape online discourse; for autonomous weapons systems; for cyber tools and autonomous cyber capabilities; to aid and make consequential decisions such as for employment, loans, and criminal sentencing; in advertising; in education and testing; in self-driving cars and navigation; in social media. Society and policy makers are rapidly trying to catch up, to adapt, to create norms and policies to guide these new areas. We see this scramble in contemporary international tax law, competition/antitrust policy, innovation policy, and national security motivated controls on trade and investment.

To understand and advise contemporary policymaking, one needs to develop expertise in specific policy areas (such as antitrust/competition policy or international security) as well as in the relevant technical aspects of AI. It is [illegible] important to build a community jointly working across these policy areas, as these phenomena interact, and are often driven by similar technical developments, involve similar tradeoffs, and benefit from similar

Long-term Risks and Opportunities

Longtermists are especially concerned with the long-term risks and opportunities from AI, and particularly existential risks, which are risks of extinction or other destruction of humanity's long-term potential (Ord 2020, 37).

Superintelligence Perspective

Many longtermists come to the field of *AI Governance* from what we can call the *superintelligence perspective*, which typically focuses on the challenge of having an AI agent with cognitive capabilities vastly superior to those of humans. Given how important intelligence is---to the solving of our global problems, to the production and allocation of wealth, and to military power---this perspective makes clear that superintelligent AI would pose profound opportunities and risks. In particular, superintelligent AI could pose a threat to human control and existence that dwarfs other natural and anthropogenic risks (for a weighing of these risks, see Toby Ord's The Precipice)[3]. Accordingly, this perspective highlights the imperative that AI be safe and aligned with human preferences/values. The field of *AI Safety* is in part motivated and organized to address this challenge. The superintelligence perspective is well developed in Nick Bostrom's *Superintelligence*, Eliezer Yudkowsky's writings (eg), Max Tegmark's Life 3.0, and Stuart Russell's *Human Compatible*. The superintelligence perspective is most illuminating under scenarios involving fast takeoff, such as via an intelligence explosion.

Problems of building safe superintelligence are made all the more difficult if the researchers, labs, companies, and countries developing advanced AI perceive themselves to be in an intense winner-take-all race with each other, since then each developer will face a strong incentive to "cut corners" so as to accelerate their development and deployment; this is part of the problem of *managing AI competition*. A subsequent governance problem concerns how the developer should institutionalize control over and share the bounty from its superintelligence; we could call this the problem of *constitution design (for superintelligence)*, since the solution amounts to a constitution over superintelligence.

Work on these problems interact. Sometimes they are substitutes: progress on managing AI competition can lower the burden on AI safety, and vice versa. Sometimes they are complements. Greater insight into the strategic risks from AI competition could help us focus our safety work. Technical advances in, say, AI verification mechanisms could facilitate global coordination (see Toward Trustworthy AI). It is imperative that we work on all promising strands, and that these fields be in conversation with each other.

Ecology and GPT Perspectives

The superintelligence perspective illuminates a sufficient condition for existential risk from AI. However, it is not necessary, and it is often the target of criticism by those who regard it as making overly strong assumptions about the character of advanced AI systems. There are other perspectives which illuminate other risks and considerations. One we might call the *AI ecology perspective*: instead of imagining just one or several systems, individually or in collaboration with humans, could give rise to cognitive capabilities in strategically important tasks that exceed what humans are otherwise capable of. Hanson's *Age of Em* describes one such world, where biological humans have been economically displaced by evolved machine agents, who exist in a Malthusian state; there was no discrete event when superintelligence took over. Drexler's *Comprehensive AI Services* offers an ecological/services perspective on the future of AI, arguing that we are more likely to see many superhuman but narrow AI services (and that this would be easier to build safely), rather than an integrated agential general superintelligence.

Another, broadly mainstream, perspective regards AI as a general purpose technology (GPT), in some ways analogous to other GPTs like steam-power, electricity, or computers (the *GPT perspective*). Here we need not emphasize only agent-like AI or powerful AI systems, but instead can examine the many ways even mundane AI could transform fundamental parameters in our social, military, economic, and political systems, from developments in sensor technology, digitally mediated behavior, and robotics. AI and associated technologies could dramatically reduce the labor share of value and increase inequality, reduce the costs of surveillance and repression by authorities, make global market structure more oligopolistic, alter the logic of the production of wealth, shift military power, and undermine nuclear stability. Of the three, this perspective is closest to that expressed by most economists and policy analysts.

These perspectives are not mutually exclusive. For example, even if we are most concerned about risks from the superintelligence perspective, the GPT perspective may be valuable for anticipating and shaping the policy, economic, and geopolitical landscape in which superintelligence would emerge.

Misuse Risks, Accident Risks, Structural Risks

Many analyses of AI risks, including many from the superintelligence perspective, understand risk primarily through the lenses of *misuse* or *accidents*. Misuse occurs when a person uses AI in an unethical manner, with the clearest cases involving malicious intent. Accidents involve unintended harms from an AI system, which in principle the developers of the system could have foreseen or prevented. Both of these kinds of risk place responsibility on a person or group who could have averted the risk through better motivation, caution, or technical competence. These lenses typically identify the opportunity for safety interventions to be *causally proximate* to the harm: right before the system is deployed or used there was an opportunity for someone to avert the disaster through better motivation or insight.

By contrast, the ecology and especially GPT perspectives illuminate a broader lens of *structural risks*. When we think about the risks arising from the combustion engine---such as urban sprawl, blitzkrieg offensive warfare, strategic bombers, and climate change---we see that it is hard to fault any one individual or group for negligence or malign intent. It is harder to see a single agent whose behavior we could change to avert the harm, or a causally proximate opportunity to intervene. Rather, we see that technology can produce social harms, or fail to have its benefits realized, because of a host of structural dynamics. The impacts from technology may be diffuse, uncertain, delayed, and hard to contract over. Existing institutions are often not suited to managing disruption and renegotiating arrangements. To govern AI well, we need the lenses of misuse risks and accident risks, but

The more we see risks from the superintelligence perspective, in which a machine agent may achieve decisive strategic advantage, especially when it emerges from rapid self-improvement beginning sometime soon, the more it makes sense to invest our attention on the cutting edge of AI and AI safety. From this perspective, the priority is to focus on those groups who are most likely to incubate superintelligence, and help them to have the best culture, organization, safety expertise, insights, and infrastructure for the process to go well.

By contrast, the more we see risks from the ecology perspective, and especially the GPT and structural risk perspectives, the more we need to understand the AI safety and governance problems in a broad way. While these perspectives may still see a comparably high level risk, that risk is distributed over a broader space of scenarios. The opportunities for reducing risk are also similarly broadly distributed. These perspectives regard it as more likely that existing social systems will be critical in shaping outcomes, important phenomena to understand, and possible vehicles for positive impact. These perspectives see greater need for collaboration, amongst a larger set of areas within AI safety and governance, as well as with experts from the broader space of social science and policymaking.

People who are foremost concerned about existential risks often prioritize the superintelligence perspective, probably because it most describes novel, concrete, and causally proximate ways that humans could lose all power (and potentially go extinct). However, the ecology and GPT perspectives are also important for understanding existential risks. In addition to illuminating other existential risks, these perspectives can illuminate *existential risk factors*[4], which are factors that indirectly affect existential risk. A risk factor can be as important to focus on as a more proximate cause: when trying to prevent cancer, investing in policies to reduce smoking can be more impactful than investments in chemotherapy.

Concrete Pathways to Existential Risk

What are some examples of concrete pathways to existential risk, or existential risk factors, that are better illuminated from the ecology and GPT perspectives?

Nuclear Instability

Relatively mundane changes in sensor technology, cyberweapons, and autonomous weapons could increase the risk of nuclear war (SIPRI 2020). To understand this requires understanding nuclear deterrence, nuclear command and control, first strike vulnerability and how it could change with AI processing of satellite imagery, undersea sensors, social network analytics, cyber surveillance and weapons, and risks of “flash” escalation of autonomous systems.

Power Transitions, Uncertainty, and Turbulence

Technology can change key parameters undergirding geopolitical bargains. Technology can lead to power transitions, which induce commitment problems that can lead to war (Powell 1999; Allison 2017). Technology can shift the offense-defense balance, which can make war more tempting or amplify fear of being attacked, disruption in relationships, gambits to seize advantage, and decline in trust. All of this can increase the risk of a systemic war, and otherwise enfeeble humanity's ability to act collectively to address global risks.

Inequality, Labor Displacement, Authoritarianism

The world could become much more unequal, undemocratic, and inhospitable to human labor, through processes catalyzed by advanced AI. These processes include global winner-take-all-markets, technological displacement of labor, and authoritarian surveillance and control. At the limit, AI could catalyze (global) robust totalitarianism. Such processes could lead to a permanent lock-in of bad values, and amplify other existential risks from a reduction in the competence of government.

Epistemic Security[5]

Arguably social media has undermined the ability of political communities to work together, making them more polarized and untethered from a foundation of agreed facts. Hostile foreign states have sought to exploit the vulnerability of mass political deliberation in democracies. While not yet possible, the spectre of mass manipulation through psychological profiling as advertised by Cambridge Analytica hovers on the horizon. A decline in the ability of the world's advanced democracies to deliberate competently would lower the chances that these countries could competently shape the development of advanced AI.

Value Erosion through Competition

A high-stakes race (for advanced AI) can dramatically worsen outcomes by making all parties more willing to cut corners in safety. This risk can be generalized. Just as a safety-performance tradeoff, in the presence of intense competition, pushes decision-makers to cut corners on safety, so can a tradeoff between any human value and competitive performance incentivize decision makers to sacrifice that value. Contemporary examples of values being eroded by global economic competition could include non-monopolistic markets, privacy, and relative equality. In the long run, competitive dynamics could lead to the proliferation of forms of life (countries, companies, autonomous AIs) which lock-in bad values. I refer to this as value erosion; Nick Bostrom discusses this in The Future of Human Evolution (2004); Paul Christiano has referred to the rise of "greedy patterns"; Hanson's Age of Em scenario involves loss of most value that is not adapted to ongoing AI market competition.[6]

Prioritization and Theory of Impact

The optimal allocation of investments (in research, policy influence, and field building) will depend on our beliefs about the nature of the problem. Given the value I see in each of the superintelligence, ecology, and GPT perspectives, and our great uncertainty about what dynamics will be most critical in the future, I believe we need *a broad and diverse portfolio*. To offer a metaphor, as a community concerned about long-term risks from advanced AI, I think we want to build a Metropolis---a hub with dense connections to the broader communities of computer science, social science, and policymaking---rather than an isolated Island.

A diverse portfolio still requires prioritization: we don't want to blindly fund and work on every problem in social more questions arise if we by default assign some weight to most areas. We thus must continue to examine and deliberate over the details of how the field of AI governance should grow.

Within any given topic area, what should our research activities look like so as to have the most positive impact? To answer this, we can adopt a simple two stage *asset-decision* model of research impact. At some point in the causal chain, impactful *decisions* will be made, be they by AI researchers, activists, public intellectuals, CEOs, generals, diplomats, or heads of state. We want our research activities to provide *assets* that will help those decisions to be made well. These assets can include: technical solutions; strategic insights; shared perception of risks; a more cooperative worldview; well-motivated and competent advisors; credibility, authority, and connections for those experts. There are different perspectives on which of these assets, and the breadth of the assets, that are worth investing in.

On the narrow end of these perspectives is what I'll call the *product model of research*, which regards the value of funding research to be primarily in producing written answers[7] to specific important questions. The product model is optimally suited for applied research with a well-defined problem. For example, support for COVID-19 vaccine research fits the product model, since it is largely driven by the foreseeable final value of research producing a usable vaccine. The product model is fairly widely held; it is perpetuated in part by researchers, who have grant incentives to tell a compelling, concrete, narrative about the value of their intended research, and whose career incentives are weighted heavily towards their research products.

I believe the product model substantially underestimates the value of research in AI safety and, especially, AI governance; I estimate that the majority (perhaps ~80%) of the value of AI governance research comes from assets other than the narrow research product.[8] Other assets include (a) bringing diverse expertise to bear on AI governance issues; (b) otherwise improving, as a byproduct of research, AI governance researchers' competence on relevant issues; (c) bestowing intellectual authority and prestige to individuals who have thoughtful perspectives on long-term risks from AI; (d) growing the field by expanding the researcher network, access to relevant talent pools, improved career-pipelines, and absorptive capacity for junior talent; and (e) screening, training, credentialing, and placing junior researchers. Let's call this broader perspective the *field building model of research*, since the majority of value from supporting research today comes from the ways it grows the field of people who care about long-term AI governance issues, and improves insight, expertise, connections, and authority within that field.[9]

Ironically, though, to achieve value in the field building model it still may be best to emphasize producing good research products. The reason is similar to that for government funding of basic research: while fellowships and grants are given primarily on the merits of the research, the policy justification typically rests on the byproduct national benefits that it produces, such as nationally available expertise, talent networks, spinoff businesses, educational and career opportunities, and national absorptive capacity for cutting edge science. I will reflect briefly on these channels of impact for AI governance, though much more could be said.

R&D, particularly in the military domain, poses substantial global risks. The space of such scenarios is vast, varying by the role and strength of governments, the nature of the risks posed by AI R&D and the perception of those risks, the control points in AI R&D, the likely trajectory of future developments in AI, and other features of geopolitics and the global landscape. I predict that any attempt to write a plan, draft a blueprint, or otherwise solve the problem, many years in advance, is almost guaranteed to fail. But that doesn't mean that the act of trying to formulate a plan---to anticipate possible complications and think through possible solutions---won't provide insight and preparation. Eisenhower's maxim resonates: "plans are useless, but planning is indispensable." To put it concretely: I believe I have learned a great deal about this problem through research on various topics, background reading, thinking, and conversations. While it is not easy for me to distill this large set of lessons into written form, I am able to mobilize and build on the most important of these lessons for any particular situation that may arise. In sum, I think there is a lot of useful work that can be done in advance, but most of the work involves us building our competence, capacity, and credibility, so that when the time comes, we are in position and ready to formulate a plan.

Consider by analogy the problem of international control of nuclear weapons. H.G. Wells imagined, in 1913, the possibility of atomic bombs, and he sketched their risks and geopolitical implications. In so doing, he helped others, like Leo Szilard, anticipate in advance (and act on) some key features of a world with nuclear weapons, such as the necessity of global control to avoid a catastrophically dangerous arms race. But in 1945-1946, actual efforts to achieve international control depended on many specific factors: agreements, misunderstandings, and conflict between the US and Soviet Union; bargains, bluffs, and brinkmanship over everything from Eastern Europe to the bomb; espionage; technical details around the control and construction of atomic weapons; allied agreements, interests, and actions; shifting opinion amongst the US public and global elites; institutional details of the UN and UNSC; business interest in nuclear energy; and the many idiosyncrasies of decision makers such as Truman, Stalin, Groves, and Baruch. Illustrating the critical role of individuals, and their beliefs and values, the most serious plan for international control---the Acheson-Lilienthal Report---wouldn't have been produced without the technical brilliance of people like Bush and Oppenheimer, was almost scuttled by Groves[10], and was ultimately distorted and poorly advocated by Baruch. Thus, even if we give ourselves the hindsight benefit of knowing the technical details of the technology, which even contemporaneous decision-makers didn't have, we see that to be able to positively intervene we would do well to have experts on-hand on a wide range of global issues, those experts should be ready to adapt their insights to the specific contours of the diplomatic problem that needs to be solved, and, lastly, those experts needs to have trusted access to those who have power "in the room".

I regard our problem as similar, but requiring an even more diversified portfolio of adaptable expertise given our greater uncertainty about technical and geopolitical parameters. Investments we make today should increase our competence in relevant domains, our capacity to grow and engage effectively, and the intellectual credibility and policy influence of competent experts.

Habiba Islam, Alex Lintz, Luke Muehlhauser, and Toby Ord.

---

[1]: "'Advanced AI' gestures towards systems substantially more capable (and dangerous) than existing (2020) systems, without necessarily invoking specific generality capabilities or otherwise as implied by concepts such as "Artificial General Intelligence" ("AGI"). AI governance definition from [www.fhi.ox.ac.uk/govaiagenda] (http://www.fhi.ox.ac.uk/govaiagenda).

[2]: Which can be defined here simply as machines capable of sophisticated information processing.

[3]: Toby Ord estimates a 1 in 10 chance of existential catastrophe from unaligned artificial intelligence within the next 100 years as compared to 1 in 10,000 from all natural risks, 1 in 1,000 from nuclear war, 1 in 1,000 from climate change, 1 in 1,000 from non-climate change mediated environmental damage, 1 in 10,000 from 'naturally' arising pandemics, 1 in 30 from engineered pandemics, 1 in 50 from other foreseen anthropogenic risks, and 1 in 30 from unforeseen anthropogenic risks. Risk from unaligned artificial intelligence thus comprises a substantial portion of Ord's total estimate of 1 in 6 for total existential risk over the next 100 years.

[4]: In Toby Ord's terminology.

[5]: I believe Shahar Avin coined this term.

[6]: Though Hanson tends to not emphasize this aspect of the scenario.

[7]: For AI safety, I would estimate there is more value in the research product, but still less than 50%.

[8]: The product model becomes more appropriate as particular governance problems come more into focus, become more urgent, and demand a written solution. At the limit, for example, would be the drafting of the constitution for an important new institution. Even in such a constitution formation scenario, however, the tacit knowledge of the involved experts continues to play a critical role.

[9]: The product model becomes more appropriate as particular governance problems come more into focus, become more urgent, and demand a written solution. At the limit, for example, would be the drafting of the constitution for an important new institution. Even in such a constitution formation scenario, however, the tacit knowledge of the involved experts continues to play a critical role.

[10]: Groves also played a huge role in promoting within U.S. decisionmakers the erroneous belief that the U.S. would retain the nuclear monopoly for a long time, an impact that was made possible by his monopoly on information about nuclear weapons and global nuclear supplies.