A reformatted copy of Allan Dafoe's Chapter 1 in The Oxford Handbook of AI Governance,
for use by FAIS Governance Autumn 2025 cohort,
by Kristian Repo

AI GOVERNANCE

Overview and Theoretical Lenses

ALLAN DAFOE

Abstract

Artificial intelligence (AI) will be a transformative technology, with extreme potential risks and benefits. *AI governance* refers to the norms and institutions shaping how AI is built and deployed, as well as the policy and research efforts to make it go well. I argue that the field of AI governance should have an expansive ambitious scope, commensurate to the challenges, with robust internal collaboration given transferable lessons and shared policy opportunities. To make sense of the impacts of AI, I offer three theoretical lenses, focusing on distinct mechanisms, impacts, and challenges. These lenses regard AI as: a general purpose technology, an information technology, and an intelligence technology. I then offer a lens on governance focusing on institutional fit and adaptation to the externalities produced by AI. Institutional adaptation will be especially difficult when a governance issue touches on deep social conflicts. Great power security competition poses a particular challenge because it can induce extreme—even existential—risks, and is among the hardest dynamics to govern. Building strong competent global institutions to govern powerful AI would be a historically unparalleled challenge, but ultimately may be required to steer away from the greatest risks inherent to great power competition.

Table of Contents

INTRODUCTION

In the coming *years*, artificial intelligence[1] will be deployed in impactful ways, such as in transportation, health, energy, news, social media, art, education, science, manufacturing, employment, surveillance, policing, and the military. As a general purpose technology (GPT) (Garfinkel, 2022; Bresnahan & Trajtenberg, 1995), the changes induced by AI will be broad, deep, and hard to foresee (Ding & Dafoe, 2021). The upsides will be substantial, but so also will be the potential disruptions and risks (Whittlestone & Clarke, 2022).

In the coming *decades*, the impacts from AI could go much further, potentially radically transforming welfare, wealth, and power to an extent greater than the nuclear revolution or industrial revolution. Speaking to this, machine learning (ML) researchers foresee the possibility of broadly human-level AI in one (8%) or two (22%) decades, and believe it more likely than not (>50%) by 2060[2]. The most rigorous attempt to date to forecast human-level AI based on mapping trends in hardware to estimates of the computational power of the brain reaches similar estimates (Cotra, 2020). The consequences will be profound, be they positive or negative.

The stakes are thus high that the development and deployment of AI go well. The field of AI governance seeks to understand and inform this challenge. To clarify I will offer some definitions. *AI governance* refers (1) descriptively to the policies, norms, laws, and institutions that shape how AI is built and deployed, and (2) normatively to the aspiration that these promote good decisions (effective, safe, inclusive, legitimate, adaptive). To be clear, governance consists of much more than acts of governments, also including behaviors, norms, and institutions emerging from all segments of society. In one formulation, the field of AI governance studies how humanity can best navigate the transition to advanced AI systems. This chapter offers a perspective on the field, emphasizing the challenges posed by significantly more advanced AI technology.

Four risk clusters

To motivate this work it can be helpful to make the potential extreme risks more concrete. Consider the following four clusters of risk, which we will discuss in more detail in the following sections.[3]

Inequality, turbulence, and authoritarianism

Declining labor share of value and a rise of winner-take-most labor markets could erode the position of labor and the relative equality underpinning democracy (Korinek & Juelfs, 2022, Boix 2022). Digitally mediated and AI filtered communication could increase polarization, epistemic balkanization, and vulnerability to manipulation, undermining liberal societies (Acemoglu, 2022). As with prior technological revolutions, these and other shocks could destabilize the social order and give rise to radical alternatives. Totalitarianism could be made more robust by ubiquitous

1 Defined here simply as machines capable of sophisticated information processing.
2 These numbers are the median forecast from a survey of ML researchers (Zhang et al., 2021).
3 This section draws from Dafoe (2020).

physical and digital surveillance, social manipulation, enhanced lie detection, and autonomous weapons.

Great-power war

Advanced AI could make crisis dynamics more complex and unpredictable, and enable faster escalation than humans could manage – a "flash war" (Scharre, 2018) – increasing the risk of inadvertent war. Advanced AI might otherwise increase the risks of war from extreme first-strike advantages, power shifts, and novel destructive capabilities (Horowitz et al., 2022).

The problems of control, alignment, and political order

AI safety is sometimes conceptualized in terms of the *control problem*, which is the problem of human intention controlling what an advanced AI does (Bostrom, 2014; Russell, 2019). An aspect of this is the *alignment problem*: constructing AI agents to have as goals the intentions of the human principal. Some experts believe controlling or aligning advanced AI systems will be difficult (Grace et al., 2018; Christian, 2020). As our AI systems increase in power, failure of control and alignment will pose ever greater risks to the users of AI and the surrounding community.

The idea of out-of-control AI systems can seem implausible to some. A reframing is to begin with the perennial problem of political order, a central challenge of which is the alignment and control of powerful social entities such as corporations, military actors, or political parties. This political "control problem" remains unsolved, in the sense that our existing solutions are patchwork and periodically fail, sometimes catastrophically with corporate malfeasance, military coups, or unaccountable political systems (Drutman, 2020). The AI control problem can be understood as analogous to the political control problem. As AI becomes more capable, autonomous, and empowering of certain social entities, these two control problems are likely to intertwine and compound.

Value erosion from competition

A high-stakes race (for advanced AI) can worsen outcomes by pushing parties to cut corners in safety. This structural risk (Zwetsloot & Dafoe, 2019) from competition can be generalized to any situation where there is a trade-off between anything of value and competitive advantage, and it can impact values beyond safety. Contemporary examples of values eroded from global economic competition could include sustainability, decentralized technological development, privacy, and equality. These *negative externalities* from competition can in principle be governed through global institutions, but adequately channeling competition can be difficult given complexity, uncertainty, rapidly evolving technology, asymmetric interests, bargaining friction, and especially great power rivalry (Coe & Vaynman, 2020; Fearon, 1995). In the long run, ungoverned military and economic competition could mean the future of humanity is pulled towards what is most adaptive within this competitive ecosystem, rather than towards what is good (for humanity, or in any other sense). Out of this anarchic competitive milieu, we might see the entrenchment and lock-in of impoverished values and forms of life (Bostrom, 2004).

Extreme risks and a holistic sensibility

Attention to the possibility of extreme and existential risks can help ensure the field invests adequately in avoiding worst case outcomes. Part of the field explicitly prioritizes attention to extreme and existential risks (as well as extreme opportunities), often theorized in terms of risks from misaligned superintelligence (Bostrom, 2014; Russell, 2019). Broadening the focus is the concept of "transformative AI" (TAI) (Gruetzemacher & Whittlestone, 2022), sometimes defined as AI which could "precipitate a transition comparable to (or more significant than) the agricultural or industrial revolution" (Karnofsky, 2016), or as AI which could lead to "radical changes in welfare, wealth, or power" (Dafoe, 2018). As the above risk clusters make clear, there are many ways that advanced AI could have extreme impacts on humanity. Analysis tends to focus on risks more than opportunities. Most believe that AI will robustly *enable* improved welfare, health, wealth, sustainability, and other social goods. Economists, for example, overwhelmingly believe AI will create benefits sufficient to make everyone better off. From this perspective, the challenge is to ensure sufficient safety and distribution of opportunity so that the benefits brought by advanced AI can be widely appreciated.

Some scholars frame different approaches as in conflict, such as between some schools of AI ethics and approaches to AI safety focused on existential risks (Piper 2022). Such a conflictual framing is unlikely to be helpful and is often misplaced (Prunkl & Whittlestone, 2020). Often the scholarship and policy work that needs to be done to address different kinds of risk overlap considerably. We see an analogous movement in AGI safety, where scholars originally prioritized thought experiments about superintelligence (Bostrom, 2014; Yudkowsky, 2008) but have increasingly built out complementary empirically informed research programs aiming for scalable advances in AI safety starting with existing systems (Amodei, et al., 2016; Hendrycks et al., 2021). Similarly, AI governance would do well to emphasize *scalable governance*: work and solutions to pressing challenges which will also be relevant to future extreme challenges. Given all this potential common interest, the field of AI governance should be inclusive to heterogenous motivations and perspectives. A *holistic* sensibility is more likely to appreciate that the missing puzzle pieces for any particular challenge could be found scattered throughout many disciplinary domains and policy areas. Overviews such as Dafoe (2018) and this *Handbook* offer a sampling of where insights might be found. We will now turn to a theoretical framework for making sense of the impacts from AI.

THEORETICAL LENSES: GENERAL PURPOSE, INFORMATION, INTELLIGENCE

How should we think about the impacts from AI? Any theoretical framework will have to balance desiderata. We would like a framework that is at a relatively high level of abstraction, so that our insights and conceptual vocabulary generalize across issue areas; however, we also want enough structure and concreteness so that it yields rich predictions. We want a framework that is

parsimonious, to be manageable; that is grounded in a compelling theoretical microfoundation and the technical features of AI; and that is close to exhaustive so as to not miss key properties.

There are many candidate properties and perspectives that we would want to highlight, such as AI as an enabling technology; the delegation of human decision-making to machines, and the encoding of politics in machines; accelerating and changing the character of decision-making processes, as well as systemic risks; accelerating economic growth, but with distributional implications; displacing labor, changing the value of capital vs labor, and impacting inequality; impacting the offense-defense balance and balance of power; altering informational dynamics like surveillance, coordination, and human imitation.

The preceding perspectives are mostly descriptions of potential impacts from AI, but they largely do not offer microfoundation for those impacts. Instead, I will offer a framework of three theoretical lenses from which these perspectives can be derived. Each of these lenses provides microfoundations and a cognate reference class, illuminating historical analogies.

These lenses are: (1) AI as a general purpose technology, (2) AI as an information technology, and (3) AI as an intelligence technology. The later categories can be understood as special cases of the earlier categories (although this conceptual nesting is imperfect). While these three lenses may seem complex or high level, I believe their richness and generality sufficiently compensates. We want a theory that not only makes sense of our present intuitions, but also allows us to anticipate and make sense of the dynamics that will later emerge. The following exposition involves many theoretical claims, concisely stated so as to sketch our current best understanding of the impacts of AI; however, these propositions can and should be questioned and studied further, and thus treated as hypotheses.

AI as a general purpose technology

We can think about AI as a *general purpose technology* (GPT). A GPT can be defined as a technology that (1) provides a valuable input to many (economic and other) processes, and (2) enables important complementary innovations (for other definitions and overviews see Garfinkel, 2022; Bresnahan, 2010; Lipsey et al., 2005). Examples include printing, steam engines, rail transport, electricity, motor vehicles, aviation, and computers; most GPTs involve either energy production, transportation, or information processing (Lipsey et al., 2005, p.133). GPTs are often attributed responsibility for long-run economic growth.

AI is a GPT, and will plausibly be the quintessential GPT. AI can serve as a fundamental input to many processes, and is highly complementary with other processes. As Kevin Kelly (2014) put it, “Everything that we formerly electrified we will now cognitize. . . . business plans of the next 10,000 startups are easy to forecast: *Take X and add AI*.” GPTs tend to be more transformative the more they are “capable of ongoing [substantial] technical improvement” (Bresnahan, 2010), which seems to be true of AI: we are still in early days of AI development, and the ceiling of potential capability likely exceeds human-level. Finally, given the plausible pace of developments, political-economic transformations from AI are likely to come more quickly than they have from most previous GPTs.

GPTs tend to have a set of important properties, which AI will likely also possess. First, GPTs grow the economy, often radically so; in fact, the concept of GPT was largely conceived to explain growth in "total factor productivity", which is a crucial component of long run economic growth. We can conceptualize this growth as arising from increases in *efficiency*, where the GPT reduces the costs of inputs to existing processes, and from *enabling* new processes altogether. The potentially Pareto-improving character of GPTs is true of AI: in principle, if deployed well and if losers are compensated, AI presents a profoundly positive opportunity for all people and groups to advance their interests, to a magnitude comparable to the industrial revolution.

However (and second), GPTs tend to be *disruptive* of existing processes, and thus also disruptive of social-political relations that depend on those processes. They tend to have substantial *distributional consequences*: shifting power and wealth, providing opportunities for certain groups, companies, and countries to rise and fall. They impose (short-run) *displacement costs* on certain groups and economic factors (e.g. land, certain kinds of capital); these costs are often not easy to identify, making it hard to insure against them or contract over them (Korinek & Stiglitz, 2019). Though the earliest versions of GPTs may appear harmless and of little utility – a cumbersome printing press; a slow prototype railway; a massive hard-to-program computer – after several generations of improvement, deployment, complementary innovations, and adaptation, their cumulative impact can be revolutionary. Even while aggregate wealth increases, some individuals, groups, countries, ideologies, and cultures will lose from these changes, if only positional goods like status. Though the net impact of the past two centuries has been favorable to labor and liberal institutions, this arguably depended on the extent to which labor and liberal institutions were (economic and military) *complements* to the new technological ecosystem, which may not continue indefinitely (see section 'AI as an intelligence technology' below).

Third, anticipation of disruption can *mobilize potential losers*, and cause social conflict. Workers, firms, and asset holders who fear being displaced may resist the technology, or seek political protections; the effects of this resistance range from minor regulatory protections to revolutions (Frey, 2019). At the international level, the (anticipated) rise and fall of countries, and the scramble for new strategic resources and capabilities, can precipitate aggressive actions and war (Horowitz et al., 2018).

Fourth, many GPTs are *strategic*, in the sense of being *essential to the military-industrial base and national power*; AI is one such strategic GPT (Ding & Dafoe, 2021). Those groups and countries that successfully harness strategic GPTs gain in relative wealth and power; in fact, possession and deployment of their mature variants is close to a necessary condition of great power status. They thus become a site of strategic investment and rivalry. GPTs often give rise to critical military assets, and are thus of interest to militaries.

Relatedly, GPTs are *dual-use*: they have both peaceful beneficial applications, and dangerous/military applications, and it is often difficult to separate these. They are sometimes developed in the military sector, sometimes the civilian, but have implications for both. For technologies with this inseparable dual-use character, arms control is especially difficult.

Each of these implications will apply to AI. By recognizing that these implications are not novel to AI, but are shared by other GPTs, we can learn from historical experience with this broader reference class.

AI as an information technology

A second theoretical lens regards AI as an *information technology.* Information is critical: for economic production, for coordination and identity, for power and bargaining, for democratic oversight and authoritarian repression. Historical information technologies have had profound impacts in generalizable ways.

An information technology is one that improves the production, compression, transmission, reproduction, enhancement, storage, control, or use of information. AI will enhance the technical possibilities for each of these processes, which will then complement the others. For example, human-machine communication will be improved through natural language understanding, bidirectional oral communication, interpretable gestures, affect and psychological inference, and contextual understanding; this will then improve the production, enhancement, and expression of information. AI assists in the *compression* of large datasets into smaller generically usable datasets, such as when converting streams of video of a pedestrian square into a digital record of who was where, when, and doing what. AI will *enhance* data by making it more searchable and readable, and by identifying useful features, which may span modalities. And of course, AI will make possible a massive amount of new uses for information, on the order of trillions of dollars' worth. AI will thus be an information technology, and will amplify other information technologies.

Some information technologies have been GPTs, inheriting the properties of GPTs discussed above. For example, speech and culture, writing, and the printing press were crucial for the rise of, respectively, homo-sapiens, civilization, and the nation-state; the telegraph and radio enabled extensive knock-on innovations, transforming war, commerce, and political order. However, information technologies have additional distinctive properties, especially if we focus on the most recent trends in digitization and digital services.

Economic implications: Increasing returns and distribution

Information technologies, and especially digital services, tend to have substantial *economies of scale*. This arises foremost because these processes involve *low marginal costs* (e.g. reproduction and transmission of information), relative to the *fixed costs* (e.g. production of a movie). A firm makes a massive ("fixed") investment to develop mapping and navigation services, and then pays a negligible cost for providing that service to the marginal user. To make this concrete for ML, the compute costs of inventing and training a new model are often orders of magnitude more than the later costs of deploying an instance of it. A second dynamic are *network economies* arising from (access restricted) communication networks: a language, telegraph network, phone network, operating system, and social network is more valuable the more other users are on it, leading to high returns to scale. These two features tend to concentrate the *global market structure* of information industries, favoring one or a few networks or firms (more on this below). A related implication is that information technologies tend to produce *winner-takes-most* labor markets, where a few superstar actors, writers, athletes, researchers, designers, entrepreneurs, and CEOs can capture most of the value in their market (Jones & Tonetti, 2020).

The preceding dynamics push toward greater *income inequality* (Korinek & Stiglitz, 2019), to firms, individuals, and even possibly countries. However, information technology has a strong

countervailing valence toward *consumption equality* because *information wants to be free*, being non-rival and hard-to-exclude. (1) Information is hard to hold on to. (i) Sometimes just the knowledge that something can be done, or the broad contours of how it is done, is a sufficient clue to dramatically accelerate a competitor's R&D catch-up. (ii) It is difficult to provide many information services without the recipient being able to copy and reproduce it, hence the elaborate (and porous) legal and hardware protections for intellectual property. (iii) The direct costs of intellectual property theft, to the thief, are often not prohibitive (as compared with other kinds of theft, such as natural resource theft); if an employee is willing to disclose information, business secrets and digital files can often be exfiltrated. (2) Ignoring the need to fund innovation, the socially efficient arrangement is to provide goods and services at their marginal cost, which in the digital realm is often close to zero. This arrangement can be achieved through public interest services (e.g. the openness norm in scientific publishing; services like Wikipedia), through limits on intellectual property (e.g. copyright limits, which enables services like Project Gutenberg), and through market competition that leads to inexpensive services (exemplified by the many free or ad-based digital services). It is hard to estimate, but plausibly the value today of free services to individuals with a smartphone is worth tens of thousands of dollars per year per person (Brynjolfsson et al., 2019). Thus, while information technologies may imbalance the income distribution, it could balance the distribution of consumer welfare. Consider a billionaire: the books, movies, video games, navigation apps, and social media services they use are largely accessible to the median wage earner.

Coordination and identity

Information technologies facilitate communication and coordination, but the political impacts are often ambiguous: innovations may strengthen or undermine existing communities and power centers.

Firstly, the economies of scale of information technologies, and complementary adaptation like standardization (e.g. in language, typography, style guides, ICANN), encourage *broader collective identities*, as information consumption can shift away from the former monopoly of local sources. This dynamic is present in the creation of national identity from more disparate local identities (Anderson, 1991), and has fueled and continues to fuel cosmopolitanism and liberalism through literature, global news, Hollywood, and the internet.

On the other hand, by allowing spatially-distributed individuals with common interests to better communicate and coordinate, information technologies may support *narrow spatially-distributed* identities, whose interests may be contrary to incumbents. Examples include global ideologies and movements (e.g. communism, environmentalism, Al-Qaeda), religions (e.g. Protestantism), and other cultural identities. Some information technologies have thus been critical in undermining existing power centers through cultural revolutions and the rise of complex spatially distributed communities. A similar proliferation of smaller organizational forms has taken place in the economy, with the information technology enabled rise of boutique firms and the gig economy.

Many of the potential impacts of AI can be interpreted through this lens of how it will structure the coordination of political communities, such as in discussions of epistemic security and the political valence of AI.

Power

Information technologies can shift power within a relationship, such as by making it easier (or harder) for one party to monitor the other or monopolize critical information. In situations of imperfect information, such as bargaining situations or principal-agent relations, becoming more informed is often critical for the distribution of value: it provides *information rents*. Information is often critical in adversarial contests, as it may help identify the plans, and physical and political vulnerabilities, of adversaries. Offering a rough proxy for the importance of information for international hard power, the US intelligence budget is 10% of its total military budget. In coup attempts, be they of the state or boardroom, "information is the greatest asset" (Luttwak, 1968, p.82), with attempts often succeeding or failing depending on the timing of when the incumbent learns about the attempt. Information is critical for domestic governance, be it effective democratic oversight or totalitarian suppression. Information technology is transforming privacy, plausibly weakening individuals' privacy against authorities, but strengthening it against social peers (Garfinkel, 2020).

The centralization of control depends on the ability of the authority to adequately monitor and communicate with its agents. The telegraph and radio dramatically curtailed the autonomy of ambassadors and ship captains. Remote and autonomous weapons will similarly empower commanders to execute orders without delegating through officers (who might object, for example, to orders to shoot civilians). Technological trends are not always towards greater centralized control, however, as exemplified by the printing press and the invention of RSA (Rivest-Shamir-Adelman) and Pretty Good Privacy encryption.

Information technologies generally increase the (economic and military) value of information and its infrastructure. This is evident in the rise of military activities in cyber and information operations via social media, and we can expect this trend to continue.

Information can move at much faster speeds than other processes: from chains of smoke signals in ancient China traveling hundreds of kilometers an hour, to the (apocryphal) use of carrier pigeons by the Rothchild's to learn of the outcome of Waterloo before others, to contemporary traders investing billions to construct inter-exchange fiber-optics and microwave beams for advantages of milliseconds. This acceleration from information-based dynamics can lead to an acceleration of crises, as exemplified by financial "Flash Crashes" where trillions of dollars in value can disappear in minutes.

The net effect of any information technology on politics and power is often hard to know in advance. It remains too early to say with confidence whether AI will strengthen the state, weaken it, or lead it to be subsumed or transformed. But it is clear that information is a critical resource for political dynamics, and AI will amplify the value and impact of that resource. These themes will recur below.

AI as an intelligence technology

The third theoretical lens involves understanding AI as a technology of *intelligence*: an innovation in the ability (of some entity) to solve cognitive tasks (Hernández-Orallo, 2017). Of the three perspectives on AI proposed here, this third is the least well developed in the literature; however, it

arguably illuminates the most important impacts of artificial intelligence. The advent of AI thus demands that we make sense of the broader reference class of intelligence and intelligence technologies. (I will only use the term technology occasionally for this lens, because it is an imperfect fit for some kinds of intelligence innovations, like the use of humans as advisors.)

Intelligence technologies can vary on a number of dimensions. One important distinction is between *tools*, on one end of the spectrum, and *systems* and *agents* on the other. Examples of tools include an abacus, a dictionary, and a notepad. These are *narrow* – they are designed to perform some specific function, and they are not meant to impact the world beyond that narrow use. These are *not autonomous* – they require a user to have impact in the world, by integrating them into some broader goal directed process. Other intelligence technologies are more general and autonomous; these, which are often the most impactful, we can call *systems* or *agents*, with "agents" denoting those that behave more like coherent goal-seeking entities. Examples of (intelligence augmenting) *systems* include the price mechanism in a free market, language, bureaucracy, peer review in science, evolved institutions like the justice system and law. Examples of *agents* are chief advisors to a monarch (e.g. the Grand Vizier), the general staff for the military, a corporation, or a deeply socialized bureaucracy. List and Pettit (2011) examine the concept of agency applied to groups. Danzig (2022) similarly analogizes AI to bureaucracies and markets, and considers with each the alignment and control problems.

We can draw out several high level properties of intelligence technologies, echoing implications we saw with general purpose technologies and information technologies. They are often critical for military and economic survival; consider the military general staff or the use of the market to allocate resources. They often transform the character of the largest political entities. Human tribes, the Neolithic state, the medieval state, and the modern state, each arose in part from improvements in intelligence technologies.

Historically, intelligence technologies both substitute for and complement (other) human cognitive labor. The use of machine calculators substituted for human "calculators", but also complemented (made more valuable) other mathematical skills. The rise of a competent bureaucracy substitutes for the actions of an individual minister, who may have formerly monopolized this aspect of policy, but also can complement a decision-maker with good judgment. An important question concerns the extent to which future AI will complement, or substitute, for human cognitive labor (Brynjolfsson, 2022), as this could have profound implications for labor share of value, inequality, and growth rates (as capital can grow itself).

Bias, alignment, and control

Perhaps, most importantly, intelligence entities often pose challenges of bias, alignment, and control.

Even simple tools can *bias* decision-making: leading us to pursue more of that which fits the tool or is made salient by the tool: *to a person with a hammer, everything looks like a nail*. Arguably, early states were biased toward legible social arrangements (Scott, 2008), contemporary policymakers focus too much on GDP (rather than actual wealth and wellbeing) (Sen et al., 2010), and social media companies optimized excessively for metrics like engagement. Because there are often political implications to any decision-making process, the introduction of cognitive tools

which shape those processes themselves have political implications. We are seeing this politics of cognitive tools in the use of AI for decision-making related to employment, crime and justice, and social relationships, but also in how tools like email and Twitter shape how people communicate, deliberate, and work.

More problematically, systems and agents may not be *aligned* with or under the *control* of the principal, such that increasing the power of the system/agent will systematically lead to outcomes in conflict with the interests of the principal. An example of system misalignment is when the market loses the ability to allocate resources well when there are significant externalities. Corporate scandals offer examples of agent misalignment, such as the Enron scandal in which Enron used complex accounting practices, and colluded with the Andersen accounting firm, to misrepresent financial performance. (Civilian) control over the military offers another prominent category of periodic agent misalignment, exemplified by the Kennedy administration's struggles with the Joint Chiefs during the Cuban Missile Crisis (Allison, 1969), or the Obama administration's struggles with US military leaders over troop requests for Afghanistan (Obama, 2020).

These problems of alignment and control can be understood as a form of Principal-Agent problem, where the agent's advantage over the principal is not just one of information but also of potentially vastly superior intelligence (Young et al., 2019). The principal may not even know what questions to ask, where to look, or have the concepts to make sense of the problem. Solutions to this problem have been explored throughout the social sciences, and in work on national governance and corporate governance, and include mechanisms for oversight, transparency, whistleblowers, representation, and other aspects of institution design. Presently, work on the problem of alignment and control for AI is almost exclusively being done by AI researchers, but given this consilience the work would benefit from experts in social science and governance (Irving & Askell, 2019).

In conclusion, it is this third lens of *intelligence* which makes clear the full extent to which AI will be transformative. Our social order depends on the alignment and control of (human and organizational) intelligence; as we augment social entities with machine intelligence, problems of alignment and control will become ever more complex and critical.

GOVERNANCE AND ANARCHY

Institutional fit and externalities

Governance involves shaping behavior to achieve social goals through institutions. "Institutions" are understood to refer to the full spectrum of social structures which shape behavior, including norms, rituals, rules, organizations, regulations, regulatory bodies, and legislatures (North, 1991).

A particularly useful conceptual tool (from economics) is that of *externalities*, which refers to byproduct impacts of an individual or group's actions on others, be they positive or negative. The central insight is that there are opportunities for institutions to increase overall welfare by discouraging behavior with negative externalities (e.g. pollution) and encouraging behavior with

positive externalities (e.g. innovations); institutions can be used to “internalize” or manage externalities.

We can then conceptualize a governance issue by examining the kinds of externalities involved – what kinds of social dilemmas emerge – to see what stakeholders, interests, and mechanisms need to be included in any institution to address them. Such a functionalist approach to institutions is common in the discipline of economics, and in rationalist approaches in political and policy sciences (e.g. Koremenos et al., 2001).

Thus, when confronting a problem of governance, we can start by asking what properties the institution will need to adequately shape behavior toward the intended social goals. What are the externalities that need to be internalized, and over what political spaces do they span? Do existing or hypothesized institutions have the needed:

- spatial remit?
- issue area remit?
- political remit, in the sense that they adequately and legitimately represent the relevant stakeholders?
- technical competence?
- institutional competence?
- influence, such as ability to sufficiently shape material incentives?

To exemplify this functionalist approach, let us consider the relatively unpoliticized issue of the governance of self-driving vehicles. The primary interests here are the safety of citizens, mobility of travelers, and business interests of vehicle producers; the primary externality is the effect of driving algorithms on other road-users. Existing traffic safety agencies have a default presumption of institutional fit for regulating this domain, given the apparent similarities to their existing remit, but we can then ask specifically how regulating self-driving vehicles might differ from regulating human-driven vehicles. We will need new methods for evaluating safety, including leveraging the opportunities from fleet crash statistics, and from evaluating algorithm performance in simulators or test environments. We have opportunities to recommend (or require) new forms of best practice, such as related to privacy, driver attention, and the storing and sharing of data from crashes. There will be implications for the pricing of insurance depending on the sharing of sensor data (when the human was driving) or the user’s choice to let the car drive. Legal institutions will need to learn how best to attribute liability between producers and users. We need to manage new risks, such as from vulnerability to hacking, and new scales of risk, such as from the possibility of a whole fleet being hacked. As an emerging industry, there may be new considerations related to supporting innovation, and thus new stakeholders and new needed technical competencies. There will be new coordination opportunities related to the building of cooperative algorithms (Dafoe et al., 2020; Dafoe et al., 2021), the setting of standards between companies, and for the creation of smart infrastructure. There will likely be trade benefits to harmonizing regulations across borders, as well international tensions around strategic barriers to trade. Some issues will grab the attention of publics and elites out of proportion to their policy importance; for example, the question of how

algorithms should ethically resolve variants of the Trolley problem* is philosophically engaging and unsettling, but largely irrelevant to the work needed to improve human well-being. Having done such an analysis of the new governance challenges, we are then in a better position to diagnose the kinds of institutions we are likely to need to well govern the domain.

The greater the distance an emerging governance area is from an existing legitimate competent institution, the greater challenges we will likely face in adapting or building adequate institutions. Some issues, though they may involve clear social benefits, can still fail if the needed institutional adaptation is just too great. As an example, consider the benefits of having every out-of-copyright book digitally available for free at local libraries; this profound public good is not being provided, not for want of a party willing to scan and provide this service, but because of congressional inability to update copyright law to legalize it (and the Department of Justice's discomfort with permitting a settlement of a class action lawsuit that would have legalized this, but only for one firm) (Somers, 2017).

Institutional adaptation or innovation will be especially difficult when the issues fall in, or are framed as part of, unresolved social conflicts. Here we may lack an overarching consensus on social goals and we may lack legitimate institutions to work with. Further, by touching on these sites of conflict the issue may itself become a battleground for the conflict, making engagement less about the issue than about the broader conflict. We could call these *deeply politicized* governance issues, understood as issues which connect to significant political conflicts at the highest levels of effective political order (i.e. usually the country). Specifically, I will reflect on the difficulty of AI governance in the presence of domestic political conflicts—between domestic political groups, and between authorities and citizens—and great power security competition.

Domestic conflicts

Many countries have significant social conflicts. For example, within the United States there is the left-right political cleavage, sometimes also referenced by the terms ”culture war” or ”polarization”. A salient example in AI governance is social media moderation, where conservatives and liberals, with different emphases, worry about censorship, bias against one's views, proliferation of “fake news”, filter bubbles, foreign intelligence operations, and mobilization of hostile social movements. Society here seems to fundamentally disagree about what it would mean to have effective, safe, legitimate social media moderation. Given deep distrust across groups about whether they have compatible social goals, it will be hard to build institutions that are widely regarded as achieving their social goals.

Another example is fairness in the use of classifiers in sensitive domains such as criminal justice, loan decisions, education, or employment. ProPublica famously sparked a discussion about racial bias in the use of algorithms for predicting the likelihood of individuals' future criminal activity, as

* The ”Trolley problem” is a set of hypothetical ethical dilemmas about sacrificing one person to save many. It reveals tensions between utilitarian and various forms of deontological and other guides to ethical behaviour. With self-driving cars it received notable attention from the ”Moral Machines experiment” which provocatively asked subjects when a self-driving car, forced to make a choice, should kill its passengers vs a set of pedestrians (of varying demographic profile). Such choices are exceedingly rare.

sometimes used in parole appeal hearings, or bail or sentencing decisions (Angwin et al., 2016). ProPublica investigated an algorithmic system commonly used in the US and reported that the false positive rate was significantly lower for white defendants than for black defendants (i.e. when certain kinds of errors were made by the algorithm, those errors tended more often to help whites and hurt blacks). Later research clarified that if any demographic differences in the three key quantities of false positive rates, false negative rates, or calibration are interpreted as evidence of "bias", then (if the classifier does not perfectly predict behavior and true crime rates differ according to the sensitive demographic trait) "bias is mathematically inevitable". By this definition, *every classifier* is "biased"; we need a more refined understanding of fairness to guide algorithmic governance.

More recent research has examined classifiers which optimize for different weightings of these quantities (Hardt et al., 2016), or indeed other relevant quantities such as conditional rates, occasionally identifying opportunities for strict improvements (Zafar et al., 2017). This case illustrates how a new capability—algorithmic decision-making—can force us to be more precise and explicit about what exactly we mean by bias, thereby opening a political debate over issues for which we lack thorough consensus, principles, and institutions (Kroll et al., 2017; Coyle & Weller, 2020). In short, one reason algorithmic bias is such a difficult governance issue is because, as a political community, we have not yet reached sufficient agreement about the principles for decision-making in sensitive domains.

AI does not just occasionally touch on pre-existing social conflicts. As smart sensors record ever more human behavior, and as decisions move from the black box of the human brain into manipulable and auditable algorithms, *AI will systematically expand the terrain for subjecting decisions to political control*. The direct effects of this can be good or bad, depending on whether political control over particular decisions would be good or bad. A systematic effect, though, is that it *increases the stakes of political contestation*. AI could thus inflame domestic political conflicts in a way analogous to how the discovery of natural resources could inflame a disputed border. We may come to regard the pre-AI era as one of significant autonomy for individuals–be they citizens, workers, and students; police officers, managers, or educators.

This systematic expansion of the possibility of political control is especially salient in conflicts between state authorities and citizens, such as over the appropriate extent of state surveillance. The existing social contract may not have been thought through and institutionalized for these new domains for state authority. Consider how Edward Snowden's leaks revealed secret US surveillance activities which had not been thoroughly publicly debated. Following the leaks, Congress did legislate and clarify the institutions around domestic surveillance. In general, when authorities and citizens in a country are at an uneasy status quo, the expansion of opportunities for political control brought by advances in AI are likely to shift the balance of power towards the state. This dynamic is seen in work on "digital authoritarianism" (Polyakova & Meserole, 2019); more work is needed to build a vigorous positive agenda of how advances in AI can strengthen liberal institutions (Hwang, 2020).

Great power security competition

The most recalcitrant and encompassing conflict is that between rival great powers—those states with the ability to exert influence on a global scale. International relations scholars characterize the enduring structural condition facing great powers as that of *anarchy*: there is no higher authority who can make and enforce laws, and so everyone's final recourse is force. Under anarchy, the shadow of power darkens all diplomacy; everything can be renegotiated, threatened, and destroyed. No one can rest secure. This vulnerability can then produce a security dilemma, where each state seeks security through their military, but in so doing makes others feel more vulnerable. The costs of this anarchy are significant: the world spends $2 trillion per year on the military (Stockholm International Peace Research Institute, n.d.), and lives with an ongoing risk of catastrophe and nuclear holocaust from thousands of nuclear warheads, 2000 of which remain on high alert (Kristensen et al., 2022). Inadequate solutions for other global public goods—such as climate change, global trade, and pandemic preparedness—are also plausibly consequences of global anarchy. To be clear, there are also significant risks from any attempted political remedy to global anarchy, namely from excessive political centralization (and, put strongly, global totalitarianism; Caplan, 2008).

AI governance, therefore, will be especially challenging for those issues that are deeply connected to great power security competition. The needed institutions for these are often global, but we may lack sufficient consensus about our social goals at that scale. Even when we have consensus (like that nuclear war is bad), the bargaining and security dynamics induced by anarchy may mean that we still cannot build the needed institutions to achieve our goals (e.g. our inability to get the world's nuclear arsenal into the small hundreds, let alone to zero).

A first cluster of governance issues in this space concerns the development of AI for lethal autonomous weapons (LAWs), cyber operations, and foreign influence operations. For each of these, but especially LAWs,[4] it is often argued that it would be desirable for countries to restrain their development and deployment of AI in certain respects. Clarifying and reinforcing norms around desirable development can shape military behavior; consider how the nuclear taboo has held back the use of nuclear weapons since 1945 (Tannenwald, 1999). However, the logic of security competition relentlessly bears down. The US Department of Defense for a long-time articulated a policy that prioritized maintaining a human-in-the-loop. However, "when instant response is imperative, even [the US] Defense Department's proponents of humans in the loop concede that their desired human control cannot be achieved" (Danzig, 2018). LaPointe and Levin (2016) conclude their article on LAWs by stating: "Military superpowers in the next century will have superior autonomous capabilities, or they will not be superpowers."

A second cluster of issues arises from the decoupling of the supply chains, commerce, and research between China and the West. China has long imposed significant constraints on Western tech companies, and has now effectively banned most Western AI services. During the past years, the US has escalated its efforts to gain independence in its supply chain for chips (such as through the CHIPS for America Act which is likely to provide ~$50 billion [Arcuri, 2022]), and ensure dependence in China's (such as by blocking export of extreme ultraviolet lithography technology from the Netherlands' ASML). Other areas of decoupling are in ML research—collaborations with

4 For example, see the Future of Life Open Letter on Autonomous Weapons (n.d.).

Chinese groups are increasingly politically scrutinized—and AI policy—consider that China is notably not included in the Global Partnership on AI, one of the most significant international initiatives on AI (Bengio and Chatila, 2021).

The AI race

An often invoked metaphor is that we are in, or entering, an *AI arms race* (Zwetsloot et al., 2018; Scharre, 2021). This metaphor is usually meant to communicate that (1) investments in AI are increasing rapidly, (2) because of a perception of a contest for very large (geopolitical) stakes, (3) and these investments are militarily relevant. This metaphor is clearly sometimes abused, such as when it is invoked to reference the "arms race" in the cost of tech talent. At present the "arms" modifier is largely literally off-point, since most of the geopolitical activity in AI is not about weapons per se, but is instead about supply chains, infrastructure, industrial base, strategic industries, scientific capability, and prestige achievements. What is not in doubt is that great powers perceive leadership in AI to be critical for future wealth and power. We might more accurately call this strategic technology competition or, more abstractly, *the AI race*.

A related metaphor, and set of ideas, is that of the race to the bottom or "race to the precipice" (Armstrong et al., 2016; Askell et al., 2019). This metaphor emphasizes how race settings —where actors perceive large gains from relative advantage—can induce actors to "cut corners", exposing the world to risks that they would otherwise prudently avoid. There are many examples of commercial races where pressure to generate profit appear to have led firms to generate excessive risks. Recent fatal AI related examples include Boeing's approach to its 737 MAX, and Uber's efforts to catch up in self-driving (although see Hunt [2020] for a positive appraisal of the strength of aviation safety institutions).

A worry is that a geopolitical race to the bottom could take place. Great powers could come to perceive themselves in a strategically decisive race for powerful AI, analogous to nuclear technology in 1945, and rush into hurried crash programs, "AI Manhattan Projects". Especially if occurring in a period of geopolitical tension, as arms races tend to, each military may be tempted to deploy powerful, but not entirely reliable, AI systems in cyber and kinetic conflict. To make the accident risks concrete, consider that in cyber-war there are likely to be significant advantages to speed, such that a human-in-the-loop would be untenable. There would likely be a possibility of unintended behavior from ML systems, especially in complex adversarial settings. Finally, there may be incentives to deploy AI-cyber systems at scale, so any unexpected extreme behavior may have broad impacts. One current aspiration is to insulate nuclear command and control from ML powered cyber-operations, so as to limit the destruction from an AI cyber accident.

How risky should we expect such a race to be? A useful starting model is the two-player strategic game known as the war of attrition, where players "bid" up the risk level or costs of conflict, until one player concedes (this is equivalent to an all pay auction, but where the "auction revenue" are the risks or costs of conflict). This is similar to the game of chicken, but with a continuous action space, and is a canonical model of actual wars of attrition and nuclear brinkmanship. Given typical simplifying assumptions such that players are rational and have common knowledge of the game, a typical result is that such two player auctions will "generate expected revenue" from each player of 1/3 of the expected value of the prize to each participant (Nisan et al., 2007, p.236). To make this

concrete, suppose that “the prize” is perceived to the decision maker as of “existential stakes”, and so similar value, relative to the status quo, as the status quo is to nuclear war; then that decision maker should be willing to race up to a 33% chance of nuclear war. A related modeling literature (Nitzan, 1994) finds that 50% of the rents are dissipated in two player rent seeking contests. So the glass is half full? The bad news is that if a prize is sufficiently attractive, such as might be perceived around a geopolitically decisive technological advantage, decision makers may be rationally willing to expose the world to a significant risk of devastation. The good news, in this model, is that these rational actors don’t race *all the way* to the bottom.

However, a real world geopolitical race may be much worse. Decision makers may not have “common knowledge” of the game, but instead may be strategically encountering it for the first time. Economists joke that if you want to quickly raise $100 dollars, host an all-pay auction for $10; invariably some participants will fail to realize that the best strategy for all but one person is to not play, and these participants will fall into an escalation spiral committing ever more funds. Further, the risks we are contemplating are novel “tail risks”, whose novelty and distance from experience will make them hard to reliably forecast. In addition, decision makers may not be rational in various senses assumed by the model: they may be overconfident, or place intrinsic value on relative outcomes or winning. When the leader of a proud country perceives honor to be at stake in a crisis, or regime survival, the costs of backing down can become much greater to them than the material issues at stake.

The above models assume that leaders can accurately perceive each other’s risky behavior. If instead risky behavior lacks a publicly observable signal or is only observed with significant noise, then it will be even harder to build viable norms and institutions for mutual restraint. The above models assume that leaders have a shared understanding of AI risk. What if, instead, AI safety is highly theory-dependent? Through a winner’s curse dynamic, and psychological and organizational rationalization, individuals and organizations may come to systematically perceive their own behavior as more safe than that of others. This can lead to an escalation spiral, where each perceives that the other is behaving much more recklessly than they are, and in turn escalates their risk taking. There are no doubt many other psychological, organizational, and political pathologies that could further exacerbate the risks from geopolitical crises.

Escaping race dynamics

How do we escape such a dangerous race dynamic? At one end of the solution space are unilateral steps to deescalate. The above models are all based on such unilateral solutions, when all racers (but one) stop racing because they perceive the risks to be excessive. Interventions that reduce the chances of underestimating risks, or that make it easier for leaders to opt of the race, therefore should be risk-reducing; however, as is often the case in models of coercion, this may also induce the other party to increase their ambition or aggression (the net effect is still usually to reduce overall risks, just by less than the direct effect, see Banks, 1990; Polachek & Xiang, 2010). Other “unilateral” solutions involve the use of force to compel an end to the race, though the possibility and execution of such options can be as risky as the race itself.

At the other end are cooperative steps to achieve *mutual restraint* (Barrett, 2007) to make the race less risky or intense. Here we seek to construct norms, treaties, and institutions to change “the rules

of the game", so that racers internalize the risks. These solutions typically require actors to reach sufficient agreement about what actions are unacceptably risky, devise means to observe compliance, and identify sufficient incentives (usually sanctions) to induce compliance. A core strategic obstacle is the transparency-security tradeoff: the arrangement must provide sufficient transparency about arming behavior to assure the other party, while minimizing the kind of transparency that compromises the security of the monitored party (Coe & Vaynman, 2020). More generally, lessons from arms control of strategic technologies such as nuclear weapons can be instructive (Maas, 2019; Zaidi & Dafoe, 2021; Scharre, 2021).

Third party organizations and global institutions may be crucial in reducing the risks from an AI race, by moving key functions of an agreement out of the halls of diplomacy, into (ideally) an impartial specialized organization devoted to the function. Trusted third parties—such as safety-focused organizations—can help articulate focal norms and standards of safe conduct. To the extent the AI race is driven by prestige motivations, as was probably true for the Space Race, third parties may be able to channel the prestige gradient towards more prosocial endeavors (Barnhart, 2022). Third party institutions may be invaluable for verifying and ruling on non-compliance, as the WTO and IAEA do in their respective areas; third party institutions can also help overcome disclosure dilemmas, enabling better sharing of information (Carnegie & Carson, 2020). Though far from our current geopolitical realities, third party institutions may even take on forms of hard power, such as imposing sanctions or directly controlling materials and activities (as was proposed for the Atomic Development Authority).

In contrast to nuclear weapons, however, military AI applications are likely to be even more difficult to control. Zaidi and Dafoe (2021) summarize:

1. **Dual use**: Powerful and dangerous AI applications will likely be harder to separate from beneficial applications, relative to nuclear technology.
2. **Value**: The economic, scientific, and other non-military value from general advances in AI will greatly exceed that from nuclear technology.
3. **Diffusion**: AI assets and innovation are much more diffused globally.
4. **Discernibility of risks**: The risks from nuclear weapons are likely easier to understand.
5. **Verification and control**: It is easier to unilaterally verify nuclear developments (nuclear tests, ICBM deployments) than deployments of dangerous cyber-weapons, and it appears easier to control key chokepoints in the development of nuclear weapons, such as with nuclear fuel and centrifuge technology. For applications of AI to cyber operations, computer hardware would be among the most tangible components, but remains deeply dual-use.
6. **Strategic gradient**: To a first approximation, the strategic value of development and innovation in nuclear weapons plateaued once a state had secure second-strike capability with thermonuclear weapons. The marginal value of the 300th warhead is small (which is why China has retained an arsenal less than 300 for nearly six decades). Decades of innovation largely haven't destabilized mutual assured destruction between the nuclear powers. AI may have a persistently steep strategic gradient, incentivizing more racing and increasing the volatility in power.

Value erosion

The discussion here, and in the literature, largely focuses on risks of catastrophic accidents. However, advanced AI and great power security competition can mix to bad effect in more subtle, gradual ways. Most of the above concerns can be theorized as arising because of a safety-performance tradeoff, and the competitive incentives that push actors to trade away safety for competitive performance. We can generalize this mechanism: any time there is a tradeoff between something of value and performance in a high stakes contest, competitive pressures can push decision makers to sacrifice that value. Contemporary examples of values being eroded by global economic competition could include optimally competitive markets, privacy, and relative equality. Mark Zuckerberg captured this logic in his prepared talking points for Congress: in response to the idea that Facebook should be broken up, Zuckerberg intended to respond that doing so would undermine a "key asset for America" and "strengthen Chinese companies" (Foroohar, 2019). In the long run, competitive dynamics could lead to the proliferation of systems (organizational types, countries, or autonomous AIs) which lock-in undesirable values. I refer to this dynamic as *value erosion*; Nick Bostrom (2004) discusses this in "The Future of Human Evolution"; Paul Christiano (2019) has referred to the rise of "greedy patterns"; Robin Hanson's (2016) *Age of Em* scenario involves loss of most value that is not adapted to ongoing AI market competition.

A common objection to the idea of value erosion is that history has seen long-term trends favoring humanity, and so empirically it does not seem like technological advances and military-economic competition lead to this kind of malign evolution. This perspective is usually coupled with a view of history as driven by the agency and intentionality of key decision makers, such as the Founding Fathers of the United States. This perspective may fail to appreciate how circumstances for humans did not obviously improve following previous technological revolutions, such as the neolithic revolution (Karnofsky, 2019). Further, the increase in human wellbeing and liberty following the industrial revolution may be attributable to the fact that human labor was made more productive, being a complement to industrial machinery, and that educated free labor became especially productive in knowledge economies. National power and wealth increasingly depended on having an educated, free, supportive citizenry. AI could change this two hundred year trend if it drastically reduces the value of human labor (by substituting more than it complements labor) and if it reduces governmental authorities' need for the support of their citizens.

Much more work is needed to understand the mechanisms and risks from value erosion. Given that value erosion operates gradually, it may be easier than acute catastrophic risks to observe and coordinate to manage. However, its slow operation and gradual entrenchment of values may also mean that sufficient attention is not mobilized in time.

Early in the evolution of flight, in the late 1920s, some military analysts came to believe that unstoppable bombers dropping poison gas over cities would fundamentally alter warfare (Zaidi, 2021, p.65). These forecasts were mistaken about the timing and the mechanism of destruction, but they correctly foresaw what would become the strategic logic of the nuclear era, made real by the discovery of the neutron chain reaction. As expressed by the title of a famous collection of essays, the early proponents of controlling nuclear weapons warned that humanity faced a choice: "One World or None". As a matter of fact, confident predictions of nuclear apocalypse were mistaken.

However, perhaps they too got the strategic logic right: increasingly powerful technology and great power competition are ultimately not compatible with the long-run flourishing of humanity.

AI governance involves building institutions to guide the development and deployment of AI to achieve our social goals. AI, however, is not a narrow technology, with limited impacts and affordances. The AI revolution will be more like the industrial revolution, transforming economics, politics, and society. To succeed, the field of AI governance must be comparably expansive and ambitious. The body of thought represented by the chapters in this *Handbook* is a great start.

ACKNOWLEDGEMENTS

I am grateful to many people for input, inspiration, and support with my thinking. For contributions relevant to this work in particular, I am grateful to: Joslyn Barnhart, Alex Belias, Nick Bostrom, Ajeya Cotra, Noemi Dreksler, Ben Garfinkel, Lewis Ho, Charlotte Jander, Ramana Kumar, Jade Leung, Tom Lue, Vishal Maini, David Manheim, Silvia Milano, Luke Muehlhauser, Toby Ord, Ken Schultz, Rohin Shah, Toby Shevlane, Robert Trager, Adrian Weller, and especially Markus Anderljung, Miles Brundage, Justin Bullock, and Anton Korinek. Thanks to Leonie Koessler and Alex Lintz for research assistance. Chapter written May 2022.

REFERENCES

Acemoglu, D., "Harms of AI" in Oxford Handbook on AI Governance, edited by Bullock, J.B., et al., Oxford: Oxford University Press, 2022, Chapter 7.5.

Allison, G.T., “Conceptual models and the Cuban missile crisis”, American Political Science Review 63, no. 3 (September 1969): 689-718. https://doi.org/10.1017/S000305540025853X.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D., “Concrete Problems in AI Safety”, arXiv, July 2016. https://arxiv.org/abs/1606.06565.

Anderson, B., Imagined Communities: Reflections on the Origin and Spread of Nationalism, London: Verso, 1991 (revised edition 2006).

Angwin, J., Larson, J., Mattu, S., and Kirchner, L., ProPublica, May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arcuri, G., “The CHIPS for America Act: Why It is Necessary and What It Does”, Center for Strategic & International Studies, January 2022. https://www.csis.org/blogs/perspectives-innovation/chips-america-act-why-it-necessary-and-what-it-does.

Armstrong, S., Bostrom, N., and Shulman, C., “Racing to the precipice: a model of artificial intelligence development", AI & Society 31, no. 2 (May 2016): 201-206. https://doi.org/10.1007/s00146-015-0590-y.

Askell, A., Brundage, M., and Hadfield, G., "The Role of Cooperation in Responsible AI Development", arXiv, July 2019. https://doi.org/10.48550/arXiv.1907.04534.

Banks, J.S., “Equilibrium Behavior in Crisis Bargaining Games”, American Journal of Political Science 34, no. 3 (August 1990): 599–614. https://doi.org/10.2307/2111390.

Barnhart, J., "Emerging Technology, Prestige Motivations and the Dynamics of International Competition" work in progress. https://www.joslynbarnhart.com/research-1.

Barrett, S., Why Cooperate?: The Incentive to Supply Global Public Goods, Oxford: Oxford University Press, 2007.

Bengio, Y., and Chatila, R., "Responsible AI Working Group Report", The Global Partnership on Artificial Intelligence, November 2021. https://gpai.ai/projects/responsible-ai/gpai-responsible-ai-wg-report-november-2021.pdf.

Boix, C. “AI and the Economic and Informational Foundations of Democracy.” in Oxford Handbook on AI Governance, edited by Bullock, J.B., et al., Oxford: Oxford University Press, 2022.

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

Bostrom, N., “The Future of Human Evolution” in Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing, edited by Tandy, C., Palo Alto, California: Ria University Press, 2004, 339-371.

Bresnahan, T., “General Purpose Technologies” in Handbook of the Economics of Innovation, Volume 2, edited by Hall, B.H., and Rosenberg, N., Amsterdam: Elsevier, 2010, 761-791.

Bresnahan, T.F., and Trajtenberg, M., “General purpose technologies ‘Engines of growth’?” Journal of Econometrics 65, no. 1 (January 1995): 83-108. https://doi.org/10.1016/0304-4076(94)01598-T.

Brynjolfsson, E., "The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence", Stanford Digital Economy Lab, January 2022. https://digitaleconomy.stanford.edu/news/the-turing-trap-the-promise-peril-of-human-like-artificial-intelligence/.

Brynjolfsson, E., Collis, A., and Eggers, F., "Using massive online choice experiments to measure changes in well-being", Proceedings of the National Academy of Sciences 116, no. 15 (April 2019): 7250-7255. https://doi.org/10.1073/pnas.1815663116.

Bullock, J.B., Huang, H., Kim, K.-C., and Young, M.M., “The Challenge of AI Governance for Public Organizations” in Oxford Handbook on AI Governance, edited by Bullock, J.B., et al., Oxford: Oxford University Press, 2022, Chapter 4.1.

Caplan, B., "The Totalitarian Threat" in Global Catastrophic Risks, edited by Bostrom, N., Cirkovic, M.M., and Rees, M.J., Oxford: Oxford University Press, 2008, 504-519.

Carnegie, A., and Carson, A., Secrets in Global Governance: Disclosure Dilemmas and the Challenge of International Cooperation in World Politics, Cambridge: Cambridge University Press, 2020.

Cellan-Jones, R., "Stephen Hawking warns artificial intelligence could end mankind", BBC, December 2014. https://www.bbc.co.uk/news/technology-30290540.

Christian, B., The Alignment Problem: How Can Machines Learn Human Values?, London: Atlantic Books, 2020.

Christiano, P., “What failure looks like”, AI Alignment Forum, March 2019. https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like.

Coe, A.J., and Vaynman, J., “Why Arms Control Is So Rare”, American Political Science Review 114, no. 2 (May 2020). https://doi.org/10.1017/S000305541900073X.

Cotra, A., “Draft report on AI timelines”, LessWrong, September 2020. https://www.lesswrong.com/posts/KrJfoZzpSDpnrv9va/draft-report-on-ai-timelines.

Coyle, D., and Weller, A., ““Explaining” machine learning reveals policy challenges”, Science 368, no. 6498 (June 2020): 1433-1434. https://doi.org/10.1126/science.aba9647.

Dafoe, A., “AI Governance: A Research Agenda”, Centre for the Governance of AI, August 2018. https://www.governance.ai/research-paper/agenda.

Dafoe, A., “AI Governance: Opportunity and Theory of Impact”, Effective Altruism Forum, September 2020. https://forum.effectivealtruism.org/posts/42reWndoTEhFqu6T8/ai-governance-opportunity-and-theory-of-impact.

Dafoe, Allan, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. "Open problems in cooperative ai." arXiv preprint arXiv:2012.08630 (2020).

Dafoe, A., Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel. "Cooperative AI: machines must learn to find common ground." Nature 593, no. 7857 (2021): 33-36.

Danzig, R., “Machines, Bureaucracies, and Markets as Artificial Intelligences”, Center for Security and Emerging Technologies, January 2022. https://cset.georgetown.edu/publication/machines-bureaucracies-and-markets-as-artificial-intelligences/.

Danzig, R., “Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority”, Center for a new American Security, May 2018. https://www.cnas.org/publications/reports/technology-roulette.

Ding, J., and Dafoe, A., “Engines of Power: Electricity, AI, and General-Purpose Military Transformations”, arXiv, June 2021. https://arxiv.org/abs/2106.04338.

Drutman, L., Breaking the Two-Party Doom Loop: The Case for Multiparty Democracy in America, New York: Oxford University Press, 2020.

Fearon, J.D., “Rationalist explanations for war”, International Organization 49, no. 3 (Summer 1995): 379-414. https://doi.org/10.1017/S0020818300033324.

Frey, Carl Benedikt. The technology trap. Princeton University Press, 2019.

Future of Life Institute, “Autonomous Weapons: An Open Letter from AI & Robotics Researchers”, no date. https://futureoflife.org/2016/02/09/open-letter-autonomous-weapons-ai-robotics/.

Garfinkel, B., “The Case for Privacy Optimism”, The best that can happen, March 2020. https://benmgarfinkel.wordpress.com/2020/03/09/privacy-optimism-2/.

Garfinkel, B., “The Impact of Artificial Intelligence: A Historical Perspective” in Oxford Handbook on AI Governance, edited by Bullock, J.B., et al., Oxford: Oxford University Press, 2022, Chapter 1.4.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., and Evans, O., "When will AI exceed human performance? Evidence from AI experts", Journal of Artificial Intelligence Research 62 (2018): 729-754. https://doi.org/10.1613/jair.1.11222.

Gruetzemacher, R., and Whittlestone, J., “The transformative potential of artificial intelligence”, Futures 135 (January 2022): 102884. https://doi.org/10.1016/j.futures.2021.102884.

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

Hardt, M., Price, E., and Srebro, N., “Equality of Opportunity in Supervised Learning”, arXiv, October 2016. https://arxiv.org/abs/1610.02413v1.

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J., "Unsolved Problems in ML Safety", arXiv, December 2021. https://doi.org/10.48550/arXiv.2109.13916.

Hernández-Orallo, J., The Measure of All Minds: Evaluating Natural and Artificial Intelligence, Cambridge: Cambridge University Press, 2017.

Horowitz, M., Pindyck, S., and Mahoney, C., “AI, the International Balance of Power, and National Security Strategy” in Oxford Handbook on AI Governance, edited by Bullock, J.B., et al., Oxford: Oxford University Press, 2022, Chapter 9.3.

Horowitz, M., Kania, E.B., Allen, G.C., and Scharre P., “Strategic competition in an era of artificial intelligence”, Center for a New American Security., July 2018. https://www.cnas.org/publications/reports/strategic-competition-in-an-era-of-artificial-intelligence.

Hunt, W., "The Flight to Safety-Critical AI: Lessons in AI Safety from the Aviation Industry", UC Berkeley Center for Long-Term Cybersecurity, August 2020. https://cltc.berkeley.edu/2020/08/11/new-report-the-flight-to-safety-critical-ai-lessons-in-ai-safety-from-the-aviation-industry/.

Hwang, T., "Shaping the Terrain of AI Competition", Center for Security and Emerging Technology, June 2020. https://cset.georgetown.edu/publication/shaping-the-terrain-of-ai-competition/.

Irving, G., and Askell, A., “AI safety needs social scientists”, Distill, February 2019. https://distill.pub/2019/safety-needs-social-scientists/.

Jones, C.I., and Tonetti, C., “Nonrivalry and the Economics of Data”, American Economic Review 110, no. 9 (September 2020): 2819–2858. https://doi.org/10.1257/aer.20191330.

Karnofsky, H., “Did life get better during the pre-industrial era? (Ehhhh)”, Cold Takes, November 2021. https://www.cold-takes.com/did-life-get-better-during-the-pre-industrial-era-ehhhh/.

Karnofsky, H., “Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity.” Open Philanthropy Project, May 2016. http://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity.

Kelly, K., “The Three Breakthroughs That Have Finally Unleashed AI on the World”, Wired, October 2014. https://www.wired.com/2014/10/future-of-artificial-intelligence/.

Koremenos, B., Lipson, C., and Snidal, D., "The rational design of international institutions", International Organization 55, no. 4 (2001): 761-799. http://www.jstor.org/stable/3078615.

Korinek, A., and Juelfs, M., “Preparing for the (Non-Existent?) Future of Work” in Oxford Handbook on AI Governance, edited by Bullock, J.B., et al., Oxford: Oxford University Press, 2022, Chapter 7.4.

Korinek, A., and Stiglitz, J., “Artificial Intelligence and Its Implications for Income Distribution and Unemployment” in The Economics of Artificial Intelligence, edited by Agrawal, A., Gans, J., and Goldfarb, A., Chicago and London: University of Chicago Press, 2019, 349-390.

Kristensen, M., Korda, M., and Norris, R., “Status of World Nuclear Forces”, Federation of American Scientists, February 2022. https://fas.org/issues/nuclear-weapons/status-world-nuclear-forces/.

Kroll, J.A., Huey, J., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G., and Yu, H., "Accountable Algorithms", University of Pennsylvania Law Review 165, no. 3 (2017): 633-706. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3.

LaPointe, C., and Levin, P.L., “Automated War: How to Think About Intelligent Autonomous Systems in the Military,” Foreign Affairs, September 2016. https://www.foreignaffairs.com/articles/2016-09-05/automated-war.

Lipsey, R.G., Carlaw K.I., and Bekar, C.T., Economic Transformations: General Purpose Technologies and Long-Term Economic Growth, Oxford: Oxford University Press, 2005.

List, C., and Pettit, P., Group Agency: The Possibility, Design, and Status of Corporate Agents, Oxford: Oxford University Press, 2011.

Luttwak, E.N., Coup d'État: A Practical Handbook, London: Allen Lane, 1968.

Maas, M.M., "How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons" Contemporary Security Policy 40, no. 3 (2019): 285-311. https://doi.org/10.1080/13523260.2019.1576464.

MacAskill, W., What We Owe the Future: A Million-Year View, London: Oneworld, 2022 (forthcoming).

MacInnes, M., Dafoe, A., and Garfinkel, B., “Anarchy as Architect: Competitive Pressure, Technology, and the Internal Structure of States”, 2022 (forthcoming).

Muehlhauser, L., “A personal take on longtermist AI governance”, Effective Altruism Forum, September 2021. https://forum.effectivealtruism.org/posts/M2SBwctwC6vBqAmZW/a-personal-take-on-longtermist-ai-governance.

Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V.V., Algorithmic Game Theory, Cambridge: Cambridge University Press, 2007.

Nitzan, S., "Modelling rent-seeking contests", European Journal of Political Economy 10, no. 1 (May 1994): 41-60. https://doi.org/10.1016/0176-2680(94)90061-2.

North, D.C., “Institutions”, Journal of Economic Perspectives 5, no. 1 (Winter 1991): 97–112. http://www.jstor.org/stable/1942704.

Obama, B., A Promised Land: Barack Obama, New York: Viking, 2020.

Ord, T., The Precipice: Existential Risk and the Future of Humanity, London: Bloomsbury, 2020.

Polachek, S., and Xiang, J., "How opportunity costs decrease the probability of war in an incomplete information game" International Organization 64, no. 1 (January 2010): 133-144. http://www.jstor.org/stable/40607983.

Polyakova, A., and Meserole, C., “Exporting digital authoritarianism: The Russian and Chinese models”, The Brookings Institution, August 2019. https://www.brookings.edu/research/exporting-digital-authoritarianism/.

Prunkl, C., and Whittlestone, J., “Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society”, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (February 2020): 138–143. https://doi.org/10.1145/3375627.3375803.

Russell, S.J., Human Compatible: Artificial Intelligence and the Problem of Control, New York: Viking, 2019.

Scharre, P., “A Million Mistakes a Second”, Foreign Policy, September 2018. https://foreignpolicy.com/2018/09/12/a-million-mistakes-a-second-future-of-war/.

Scharre, P., "Debunking the AI Arms Race Theory", Texas National Security Review 4, no. 3 (Summer 2021): 121–132. http://dx.doi.org/10.26153/tsw/13985.

Scott, J.C., Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed, New Haven: Yale University Press, 2008.

Sen, A., Fitoussi, J.P., and Stiglitz, J., Mismeasuring Our Lives: Why GDP Doesn't Add Up, New York: The New Press, 2010.

Somers, J., “Torching the Modern-Day Library of Alexandria”, The Atlantic, April 2017. https://www.theatlantic.com/technology/archive/2017/04/the-tragedy-of-google-books/523320/.

Stockholm International Peace Research Institute, “SIPRI Military Expenditure Database”, no date. https://www.sipri.org/databases/milex.

Tannenwald, N., “The Nuclear Taboo: The United States and the Normative Basis of Nuclear Non-Use”, International Organization 53, no. 3 (Summer 1999): 433–468. https://doi.org/10.1162/002081899550959.

The Initiative on Global Markets, “Robots and Artificial Intelligence”, September 2017. https://www.igmchicago.org/surveys/robots-and-artificial-intelligence-2/.

Whittlestone, J. and Clarke, S. “AI Challenges for Society and Ethics” in Oxford Handbook on AI Governance, edited by Bullock, J.B., et al., Oxford: Oxford University Press, 2022.

Young, M.M., Himmelreich, J., Bullock, J.B., and Kim, K.-C., "Artificial Intelligence and Administrative Evil", Perspectives on Public Management and Governance 4, no. 3 (September 2019): 244-258. https://doi.org/10.1093/ppmgov/gvab006.

Yudkowsky, E., “Artificial Intelligence as a Positive and Negative Factor in Global Risk”, in Global Catastrophic Risks, edited by Bostrom, N., and Ćirković, M.M., Oxford: Oxford University Press, 2008, 308–345.

Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P., and Weller, A., “From Parity to Preference-based Notions of Fairness in Classification”, arXiv, November 2017. https://arxiv.org/abs/1707.00010.

Zaidi, W.H., Technological Internationalism and World Order Aviation, Atomic Energy, and the Search for International Peace, 1920–1950, Cambridge: Cambridge University Press, 2021.

Zaidi, W., and Dafoe, A., "International Control of Powerful Technology: Lessons from the Baruch Plan for Nuclear Weapons", Centre for the Governance of AI, March 2021. https://www.governance.ai/research-paper/international-control-of-powerful-technology-lessons-from-the-baruch-plan-for-nuclear-weapons.

Zhang, B., Dreksler, N., Anderljung, M., Kahn, L., Giattino, C., Dafoe, A., and Horowitz, M.C., “Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers”, retrieved November 2021 from https://osf.io/v7f6g/.

Zwetsloot, R., and Dafoe, A., “Thinking About Risks From AI: Accidents, Misuse and Structure” Lawfare, February 2019. https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure.

Zwetsloot, R., Toner, H., and Ding, J., “Beyond the AI Arms Race: America, China, and the Dangers of Zero-Sum Thinking”, Foreign Affairs, November 2018. https://www.foreignaffairs.com/reviews/review-essay/2018-11-16/beyond-ai-arms-race.