Annual Review of Political Science

Terra Incognita: The Governance of Artificial Intelligence in Global Perspective

Allison Stanger,[1] Jakub Kraus,[2] Woojin Lim,[3] Georgia Millman-Perlah,[1] and Mitchell Schroeder[4]

[1]Department of Political Science, Middlebury College, Middlebury, Vermont, USA; email: stanger@middlebury.edu, gmillmanperlah@middlebury.edu

[2]Center for AI Policy, Washington, DC, USA; email: jakub@aipolicy.us

[3]Department of Philosophy and Government, Harvard University, Cambridge, Massachusetts, USA; email: woojin_lim@protonmail.com

[4]Middlebury Institute of International Studies at Monterey, Monterey, California, USA; email: mschroeder@middlebury.edu

First published as a Review in Advance on April 15, 2024

The *Annual Review of Political Science* is online at polisci.annualreviews.org

https://doi.org/10.1146/annurev-polisci-041322-042247

Keywords

artificial intelligence, AI, open societies, sustainability, governance, plurality, democracy, inequality

Abstract

While generative AI shares some similarities with previous technological breakthroughs, it also raises unique challenges for containing social and economic harms. State approaches to AI governance vary; some lay a foundation for transnational governance whereas others do not. We consider some technical dimensions of AI safety in both open and closed systems, as well as the ideas that are presently percolating to safeguard their future development. Examining initiatives for the global community and for the coalition of open societies, we argue for building a dual-track interactive strategy for containing AI's potentially nightmarish unintended consequences. We conclude that AI safety is AI governance, which means that pluralist efforts to bridge gaps between theory and practice and the STEM–humanities divide are critical for democratic sustainability.

INTRODUCTION

In the June 1955 issue of *Fortune*, the polymath genius John von Neumann, an intellectual bridge between the atomic and digital epochs, published a remarkable meditation on the existential threats facing humanity. Von Neumann's essay, "Can We Survive Technology?," begins with a dire warning: "literally and figuratively, we are running out of room" (von Neumann 1963, p. 505). Von Neumann was of course speaking of the nuclear age, but the same could be said of the rapid developments in artificial intelligence (AI) today (Ord 2022). In both instances, the velocity of human innovation has outstripped the human capacity to oversee it, threatening human values.

Despite those striking similarities, the nuclear and AI revolutions are different in kind, especially from an epistemological perspective. The creators of ChatGPT and its subsequent iterations do not understand exactly how it arrives at its outputs, which suggests they cannot fully predict its future behavior (Ganguli et al. 2022, Wei et al. 2022, Schaeffer et al. 2023). As AI models have grown more capable and more broadly deployed, our understanding of how they work internally remains limited. For example, it might be difficult to detect from their outputs whether they use biased heuristics or engage in deception (Park et al. 2023). The emergent and general-purpose qualities of large language models (LLMs) have increased the uncertainty surrounding their future effects.

The good news is that over the summer of 2023, the world at large finally woke up to the wolf at its doorstep, and a plethora of initiatives have been designed to provide guardrails and moats for technological innovation as it accelerates. In April 2023, China published "Measures for the Management of Generative Artificial Intelligence Services," which came into effect in August 2023 (Sheehan 2023). In May 2023, the G7 launched the disturbingly named Hiroshima AI Process. The following month, the European Parliament endorsed the draft EU AI Act. In July 2023, UN Secretary-General António Guterres called for the creation of a global AI regulatory watchdog. The US Congress has a wide range of bills in process, and in August 2023, the Biden administration, under the leadership of DARPA (the Defense Advanced Research Projects Agency) and in collaboration with Google, Microsoft, OpenAI, and Anthropic, launched the AI Cyber Challenge (White House 2023a). Clearly, this is a rapidly evolving subject.

Writing a review of existing literature is therefore a bit like attempting to drink from a firehose. Our aim is to describe the lay of the land as of mid-December 2023, suggest some new areas for future research, and frame the entire endeavor in a way that lends greater coherence to a growing body of work in progress, while giving some perspective on how one might situate future new developments as they unfold. Throughout, we deploy Dafoe's (2022, p. 1) definition of AI governance as "the norms and institutions shaping how AI is built and deployed, as well as the policy and research efforts to make it go well."

For general orientation purposes, the history of cyberspace to date encompasses three distinct stages. Web2 refers to the social internet, which was built upon the original read/write-only blogosphere internet (Web1). Web3 refers to the internet of decentralized AI, the metaverse, and blockchain-inspired technologies from crypto to decentralized autonomous organizations such as BlueSky and Mastodon.

The article starts by arguing that the challenge of governing artificial intelligence is more aptly described as one of governing *alien* intelligence, a powerful engine of revolutionary change. LLMs, such as ChatGPT, are triumphs of emergence, generating results that cannot be explained with Newtonian precision (Kauffman 2008). While generative AI shares some similarities with previous technological breakthroughs, it also raises new challenges for containing its negative social and economic consequences. We highlight states' approaches to AI governance within their polities and investigate how they do or do not build a reasonable foundation for transnational governance.

We then consider the technical dimensions of AI safety in both open and closed systems and the ideas that are presently percolating to safeguard national security. Next, we examine initiatives for the global community and for the community of open societies, arguing for the need to distinguish between the two and to build a dual-track strategy in order to harness AI's potential for good. The concluding section gives special attention to principles for effective AI governance and cross-regional opportunities for collaboration, which provide additional insurance against the negative consequences of the accelerating velocity of change humanity now faces.

HOW AI GOVERNANCE DIFFERS FROM PREVIOUS GOVERNANCE CHALLENGES

The modern world is full of complex machines that users rely on experts to understand; if my car breaks down, I do not know how to fix it, but a mechanic probably can. Generative AI is something else entirely. No single human understands precisely how it arrives at its responses, at least not in the sense of an Enlightenment-style cause-and-effect sequence. If it goes off the rails, as it notably did when it tried to persuade *New York Times* reporter Kevin Roose that he should divorce his wife and instead devote himself to his new silicon friend, a band-aid such as reinforcement learning from human feedback (RLHF) can be applied to the model to prevent similar behavior in the future. RLHF relies on human ghost work, the invisible behind-the-scenes labor whereby human grooming of raw data makes generative AI possible (Gray & Suri 2019). However, that RLHF adjustment does not happen instantaneously; the model needs to be retrained to behave differently in the future, which takes time.

Additionally, there are limitations to the safety benefits of RLHF, and with the advent of open-source foundation models that can run on a laptop, some future users might intentionally hijack ostensibly benign models and turn them to malevolent purposes (Casper et al. 2023). By a foundation model, we mean any model that is trained on large amounts of data and can be adapted via tuning to a wide range of downstream tasks.

Because the costs of building state-of-the-art foundation models are so high, only wealthy companies can muster the resources to bring them into being (Knight 2023, Maslej et al. 2023). It would thus seem that foundation models confer an enormous advantage on Big Tech, but it is one that open access could disrupt. In March 2023, for example, Meta released an open-access foundation model, ostensibly to challenge the predominant closed model at the time, ChatGPT. This open-source rival prompted one Google employee to conclude, in a leaked memo that circulated widely (Patel & Ahmad 2023), "we aren't prepared to win this arms race and neither is OpenAI." Microsoft and Meta together introduced the more powerful open-access Llama 2 in July 2023. Its power rivals that of closed models, yet it can be easily customized at minimal incremental cost by one human with a laptop and a free evening (Touvron et al. 2023).

In its capabilities, GPT4 was a great leap forward from its predecessor GPT3.5, and that massive improvement took place in a year's time. If more powerful versions of the model were to deliver deadlier ideas (such as a do-it-yourself, step-by-step recipe for a lethal new pathogen) that spread virally, as everything does in today's hyperconnected world, great irreparable damage will be done before a course correction might be made (Soice et al. 2023).

Although there are thoughtful scholars who might disagree (Mitchell 2020, 2023), we see ourselves poised at the dawn of a new era, especially concerning challenges for governance in an environment of unprecedented corporate power. Foundation models confer enormous additional relative power on Big Tech vis-à-vis national governments. In the September/October 2023 issue of *Foreign Affairs*, Mustafa Suleyman, a cofounder of Google DeepMind and now the CEO of Inflection (an AI company), joined with political scientist Ian Bremmer to argue that the time has come

to invite technology companies to the international governance table. They note that these actors "may not derive legitimacy from a social contract, democracy, or the provision of public goods, but without them, effective AI governance will not stand a chance" (Bremmer & Suleyman 2023).

This prescription boils down to letting large, American-headquartered technology companies govern themselves while amassing astonishing wealth. Multinational tech giants in the West have become so powerful that they function like transnational proto-governments, shaping the context for federal governance in profound ways (Lanier & Weyl 2018). While Bremmer and Suleyman surely see themselves as pragmatists, to many American voters their argument sounds like an endorsement of techno-plutocracy.

Bremmer and Suleyman go on to propose three overlapping governance regimes: "one for establishing facts and advising governments on the risks posed by AI, one for preventing an all-out arms race between them, and one for managing the disruptive forces of a technology unlike anything the world has seen." Our article surveys the existing literature and proposes potential solutions through a different tripartite scheme, as we take our point of departure from theoretical understandings of governance and power, rather than those of technology and power.

First, there are the initiatives that can be taken at the domestic (national or subnational) level. Second, there is the matter of global governance, which can span both closed and open societies, when possible, or focus on solutions that the democratic world alone can undertake collectively to render democracy sustainable. Finally, there is the matter of technical AI safety, or safeguards that can be baked into the technology itself. Both large and small technology companies, as well as academia, have a significant role to play in technical AI safety, whereas the first two categories are the rightful domain of governments and civil society networks. At each of these levels are novel and pervasive risks, ones that existing governance structures were not designed to contain.

Along parallel lines, the US Constitution's Bill of Rights was originally crafted to protect citizens from government encroachments on their privacy or freedom of speech and assembly, but in today's virtual and privatized public square it is social media companies that decide when freedom of expression prevails and when censorship carries the day. Section 230 of the 1996 Communications Decency Act enables Web2 companies to monitor speech and user content as they see fit, with the speaker held responsible for harmful speech, not the service that hosts and amplifies it. Platforms have done so to date in a way that serves their bottom line rather than any rule set, either implicit or explicit, designed to mediate between competing values (Lanier & Stanger 2024). As Lessig (2006, p. 236) foresaw nearly two decades ago, "the architecture of cyberspace is the real protector of speech there; it is 'the real First Amendment in cyberspace,' and this First Amendment is no local ordinance."

The failures of current internet and data governance have highlighted the trade-offs inevitably involved in preserving privacy and protecting national security simultaneously (Stanger 2019). The ad-driven business model that powers Meta and Google, which turns personal data into a new asset class that is often exploited without users' content, is what Zuboff (2019) has called "surveillance capitalism." The digital ecosystem these Web2 forces created was ripe for foreign electoral interference. At the time of this writing, all the prospects for voter manipulation we witnessed in the 2016 and 2020 US elections remain, and with the addition of automated misinformation via generative AI, electoral interference may be a serious challenge for the range of important elections on the horizon in 2024.

While the nuclear arms race inspired collaborative efforts at arms control that merit replication whenever possible, critical differences between the likely impact of the two technological breakthroughs exist. First, since it is, in the end, just software, AI can proliferate rapidly, whereas the International Atomic Energy Agency's oversight of enriched uranium presents a significant (although not insurmountable) obstacle to nuclear proliferation. Second, there are a limited number

of direct applications of nuclear technology, such as weapons and bombs, whereas AI is a general-purpose technology that may affect nearly every economic sector as well as having a range of military applications. Finally, AI weapons systems present the possibility of inanimate systems themselves wielding massively destructive force. It might start with AI-operated wingmen in the Air Force (Lipton 2023), but the possibility of AI-controlled drone armies is no longer beyond reach (Russell 2022, Anderljung & Hazell 2023).

The interdependent global financial system, which is driven by algorithms, has produced global financial crises that have yet to be properly addressed with global governance to prevent their recurrence (Stiglitz 2002). The Dodd-Frank financial reforms after the 2008 crash were designed to minimize the probability of a repeat cascade of disastrous effects. These reforms were compromised in the years that followed. The details are unimportant for our purposes here, but we would point out that automated trading based on breakthroughs in generative AI will only further turbocharge the velocity of transactions, increasing the prospects for emergent phenomena that could bring down the entire complex, adaptive system.

While there are historical instances of profound technological change and governance responses designed to meet them, those problems are dwarfed by complicated ramifications of an interdependent world of unbound generative AI. The increased velocity of technological change (Kurzweil 2006, Amodei 2023), the intangible nature of AI assets compared to other assets (Harari 2017) and the globalized nature of AI research and development (Susskind & Susskind 2022) place us in uncharted territory.

Frontier AI, defined as "highly capable models which could have dangerous capabilities that are sufficient to pose severe risks to public safety" (Anderljung et al. 2023, p. 2; see also Shevlane et al. 2023), is a useful category for referring to post-GPT4 types of generative AI we are likely to see in the next year as Silicon Valley races to scale existing models with previously unfathomable amounts of computational power (compute). The concept of alien intelligence takes on new salience as one contemplates machine models that can execute internet tasks and communicate with one another with minimal human intervention. Going further, hypothesized superintelligent machines could very well develop new concepts and discoveries that could be incomprehensible to us, in the same way that an iPhone or automobile would be bewildering to Ancient Greek philosophers.

Humanity is at a crossroads. As German philosopher Martin Heidegger put it in his prescient 1954 essay, "The Question Concerning Technology," the individual stands at "the very brink of a precipitous fall; that is, he comes to the point where he himself will have to be taken as standing-reserve" (Heidegger 1977, p. 27). Humans may be on the verge of being eclipsed by their own creations, which they themselves do not fully understand, and the epistemological gap between human and machine, absent human intervention, is likely to grow. The promise and peril of frontier AI "renders humanity a very small phenomenon compared to something else that is far more intelligent and will become incomprehensible to us," says humanist Douglas Hofstadter, author of *Gödel, Escher, Bach*, "as incomprehensible to us as we are to cockroaches" (quoted by gwern 2023).

STATE APPROACHES TO AI GOVERNANCE

New AI capabilities will have vast consequences for national security, global economic stability, and planetary sustainability. The world's predigital legacy governance institutions are designed for a world that is no more; technological innovation has outstripped the requisite legal and institutional frameworks to channel it in the public interest. Code is law (Lessig 2006), and code is proliferating quickly in both open-source and closed systems.

In the United States, the Biden administration has sought to confront the challenge by promoting AI safety within tech companies and voluntary restraint among Big Tech's major players

(White House 2023d). Its November 2023 Executive Order aspires to a rights-based approach, but the devil will be in the details (White House 2023c). Various legislative efforts have been introduced in Congress but, to date, have produced no significant regulations of industry. The American regulatory apparatus, in practice, has followed a largely market-driven approach. There are voluntary standards on the books and an emerging attempt to create legal and regulatory standards.

An EU regulatory apparatus, in contrast, already exists and focuses on protecting rights. The 2018 General Data Protection Regulation (GDPR)[1] protects the privacy and personal data of EU citizens; the 2022 Digital Markets Act[2] ensures fair and open digital markets by targeting the large "gatekeeper" companies that control access; and the new 2023 Digital Services Act[3] regulates online platforms, especially big ones, to ensure a safer and more open digital space. These are all examples of the rights-based regulatory approach in practice, and the new 2023 AI Act does not deviate from the general template (Hoffman 2023).

What Bradford (2020) has aptly called "the Brussels Effect" occurs when the European Union regulates the global market indirectly by imposing its standards and regulations on European member states and multinational corporations, which then has ripple effects beyond Europe's borders. For example, the European Union has insisted that Apple have one port for all its future products, instead of iteratively deploying new ports for the latest model, and Apple has announced that it will likely apply the EU ruling to the American market as well, responding to the long-standing complaints of Mac users who had been perpetually required to update their peripherals. The new iPhone 15 features the more general USB-C port rather than the iPhone 14's eccentric lightning port. The Brussels Effect has caused some to complain that the United States is exporting its regulatory responsibilities to Europe.

In *Digital Empires: The Global Battle to Regulate Technology*, Bradford (2023a) compares "the Chinese state-driven regulatory model" with the US market-driven and European rights-driven approaches. She argues that the European rights-focused model is better equipped to unite the free world as a counterpoint to China's growing power in the digital realm. The European Union has been both a trailblazer and a unique real-time laboratory for exploring the intended and unintended consequences of efforts to harness technology in service of public goods (Bradford 2023b).

The Chinese state-driven regulatory model is, in reality, the Chinese Communist Party–driven regulatory model. Every Chinese state institution or company has a related Party cell within it to ensure that every decision aligns with Party values (Pearson et al. 2022). As such, the state-driven regulatory model does not require a robust audit function since it is effectively built into the very bones of Chinese organizations and civil society (Strittmatter 2020).

In contrast, EU law's principle of subsidiarity, which ensures that decisions are made as closely as possible to the citizen, mediates the relationship between European, national, and local laws. Subsidiarity ensures that the European Union acts only when necessary, respecting the diversity of its member states. Specifically, the European Union should not take legislative action unless it is demonstrably more effective than action taken at the national, regional, or local level (European Parliament 2024). Upholding the principle of subsidiarity thus indirectly serves an auditing or oversight function.

[1] See **https://gdpr.eu/what-is-gdpr/**.
[2] See **https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en**.
[3] See **https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en**.

The American approach, with its emphasis on promoting innovation and letting market mechanisms rule whenever possible, not only has been short on producing regulation but also has essentially no oversight mechanism, save constitutional review, to align the digital sphere with either user rights or citizen needs. The Facebook Global Oversight Board is a good example of a self-imposed industry effort to secure some global accountability. It was intended to oversee Facebook's content moderation decisions, but in the end, Chief Executive Officer Mark Zuckerberg did not deliver on his promises, and the standard American laissez-faire approach of minimal government intervention prevailed (Stanger 2024). In comparison with Europe and China, the tech industry has largely called the shots to date in the United States, and there has been no significant new legislation targeting the adverse public consequences of the ad-driven business model since the 2016 elections debacle.

One might at first think that these differing regulatory models would simply produce different digital ecosystems that in turn define digital sovereignty (a country's control over its data and technologies) differently in the United States, European Union, and China, but that would be to overlook the transnational dimensions of the data flows themselves. In reality, the movement of data from large technology companies in the United States to China threatens digital sovereignty around the world. In prioritizing growth and innovation, the US regulators have overlooked predatory data-gathering practices. The Chinese government profits from US-based technology companies' practice of turning users into the product (when the product is free, so the saying goes, you are the product) and turning personal data into profit. When these same companies seek access to the Chinese market, they amplify China's national power. "Together, the US and Chinese approaches to data security increase the odds that citizens' data will be moved across international borders without those citizens' consent" (Kokas 2023, p. 2).

Frontier models present a distinctive regulatory challenge, which requires a tripartite approach to framing appropriate regulation. First, multi-stakeholder standard-setting conversations must be set in motion to identify the necessary requirements for frontier AI developers. Second, registration and reporting requirements must be outlined to allow regulators to oversee ongoing AI development processes. Finally, any effective regulatory framework must articulate mechanisms of compliance with technical AI safety standards. One way of spanning these three building blocks for regulating frontier models is to require predeployment risk assessments that would be scrutinized by external bodies to encourage their consideration in deployment decisions. Yet, the rapid pace of development demands a nimble approach, as continual response to new information about capabilities postdeployment will be required (Anderljung et al. 2023). It will not be enough to clear models for deployment, as frontier models will continue to learn and manifest new capabilities after product releases.

The three necessary conditions for trustworthy AI all depend on greater transparency at every stage of the development and deployment process than presently exists. A recent Stanford study evaluated whether the most prominent foundation models comply with the draft requirements of the AI Act and found that they do not. "Foundation model providers rarely disclose adequate information regarding the data, compute, and deployment of their models as well as the key characteristics of the models themselves. In particular, foundation model providers generally do not comply with draft requirements to describe the use of copyrighted training data, the hardware used, and emissions produced in training, and how they evaluate and test models" (Bommasani et al. 2023a). In short, "pervasive opacity compromises accountability for foundation models" (Bommasani et al. 2023b). Yet, public transparency would potentially undermine a US national security advantage. The trade-offs between security and openness should not be underestimated.

In its Blueprint for the AI Bill of Rights, the Biden administration set up five areas as a guide for society to avoid the threats that AI poses to the nation: "safe and effective systems, algorithmic

discrimination protections, data privacy, notice and explanation, and human alternatives, considerations, and fallbacks" (White House 2023b). In practice, however, prior to the release of the November 2023 Executive Order (White House 2023c), the Biden White House had primarily focused on organizing industry self-regulation with government audits in hopes of fueling innovation to keep America's national security advantage strong (White House 2023d). Since frontier AI models are largely products of US-headquartered companies, this practice is a green light for innovation. However, it also partly explains why the results are not EU AI Act–compliant, nor are they likely to be, absent a more concerted effort at US–European coordination.

There have been repeated attempts to pursue antitrust prosecution against Big Tech, with Amazon in the crosshairs of Federal Trade Commission Chair Lina Khan (Khan 2017) and strong criticism from both left (Teachout 2020) and right (Hawley 2021) on the erosion of citizen rights through the profiteering of technology companies. Senators Elizabeth Warren (D-MA) and Lindsey Graham (R-SC) have proposed bipartisan legislation for a new American agency to rein in Big Tech (Graham & Warren 2023). The market-driven US regulatory response to date may well be challenged in the runup to the 2024 elections, as Americans become more aware of the malevolence that frontier models can unleash.

Since LLMs are fed with human-created materials, the automation of human bias poses a significant challenge for democratic fairness, especially in the presence of unjust institutions, where inequality before the law can be camouflaged by algorithmic precision. If the model learns from biased data, it may perpetuate injustice rather than challenging it (Noble 2018, O'Neil 2016). If institutional racism contributes to inequality before the law, then making the system fair will not ensure that it is just. Since fairness has no value in the absence of just institutions, better governance of AI for justice and fairness might (*a*) adopt a values-first approach to bias interventions, (*b*) decouple decisions processes, (*c*) explicitly model social injustice, (*d*) intervene to increase data quality, and (*e*) use weighted lotteries (Mullainathan & Obermeyer 2021, Vredenburgh 2022).

Discrimination, violations of privacy, and other forms of algorithmic unfairness challenge our ability to realize rights-based ideals, especially in rapidly changing technological circumstances. A particular area of concern is ensuring that new AI systems do not simply reify existing prejudices, favoring certain societal groups while discriminating against others (Clement-Jones 2023). "For some challenges, received policy instruments may be sufficient. But for the most significant ones, democratic politics need to develop and implement genuinely new forms of collective political action to control where and how AI is implemented" (Viehoff 2022, p. 1). AI presents possibilities for enlarging the communities from which the original legitimizing social contract was constructed. Put another way, we currently have no global AI governance, but many of the most pressing challenges are global in scope, as a global community emerges in search of new institutional forms.

The current generation of generative AI is best understood as a force exacerbating and accelerating existing trends in domestic and international politics (Arsenault & Kreps 2022), rather than one that breaks dramatically with the politics of the past. Thus, the divide between authoritarianism and liberal democracy is salient, as is the rising collective power of Brazil, Russia, India, China, and South Africa, which appear to be functioning as the contemporary variant of the Cold War nonaligned movement (Larsen 2022).

AI SAFETY IN CLOSED AND OPEN SYSTEMS

AI safety and AI governance share common aspirations for developing technology that benefits humanity rather than diminishing us. "AI safety focuses on the technical questions of how AI is built; AI governance focuses on the institutions and contexts in which AI is built and used" (Dafoe

2018, p. 6). Frontier AI is likely to bring with it a range of new dangers: AI-powered cyberattacks, automated propaganda, turbocharged political strategy, bioweapons of mass destruction, unaligned spontaneous action, and automatic viral propagation of all the above (Shevlane et al. 2023). Many of these security threats are less easily contained in open societies than in techno-authoritarian ones, which poses existential risks for both plurality and liberal democracy (Allen et al. 2024).

Advances in generative AI promise unprecedented benefits for humanity, but the very generativity that makes the spectacular advances possible also presents tremendous societal risks that require proactive management. The rapid and massive scaling of existing models presents risks along multiple dimensions. First, dangerous capabilities can emerge spontaneously, so past remedies are no guarantee of preventing future catastrophe. Put another way, it is challenging to predict what brilliant insight with malevolent undertones may arise and when. Second, the velocity of change makes both closed and open-source models difficult to regulate. Further, "Due to the considerable power imbalance between users of AI in comparison to those AI systems are used on, successful regulation might be difficult to create and enforce. As such, AI regulation is more of a political and socio-economic problem than a technical one" (Weissinger 2022, p. 1).

AI safety requires addressing the challenge of regulating closed-source frontier AI models through public–private standard-setting, as well as registration and reporting requirements that provide free-world regulators with a window on development processes. In addition, mechanisms for ensuring compliance with safety standards for both development and deployment of frontier AI models must be established. A baseline set of safety standards might include "conducting predeployment risk assessments; external scrutiny of model behavior; using risk assessments to inform deployment decisions; and monitoring and responding to new information about model capabilities and uses postdeployment" (Anderljung et al. 2023, p. 1).

The question of standards is a global one. Faced with Europe's rights-based approach, the United States' and United Kingdom's hands-off orientation, and China's required adherence to the "core values of socialism" (Espinoza et al. 2023), companies may vote with their feet and deploy their latest closed-source inventions only in friendly ecosystems. As a promising development, four of the most influential AI companies—Anthropic, Google, Microsoft, and OpenAI—joined forces in July 2023 to form the Frontier Model Forum, which will be used to cooperate on responsible development of AI (Microsoft 2023).

Senator Chuck Schumer (D-NY) summoned all the big players to Washington, DC, in September 2023 to have a hand in shaping new, bipartisan AI regulation. Critics pointed out that the last round of pledged public–private cooperation over abuses of social media resulted only in the perpetuation of the status quo. Reining in Big Tech's power has been a Washington agenda item for more than five years, but to date Congress has not passed "a single comprehensive law to protect data privacy, regulate social media or promote fair competition between tech giants," despite numerous congressional hearings on these topics (Espinoza et al. 2023). The US National Institute of Standards and Technology (NIST) has developed the Artificial Intelligence Risk Management Framework (AI RMF) for voluntary AI risk management, but even NIST itself acknowledges that there is still a need to translate their high-level principles into practice (Barrett et al. 2023). Stanford's Center for Advanced Study in the Behavioral Sciences responded to that call with a field-building effort to translate theory into practice (Guszcza et al. 2022).

The bipartisan framework for AI proposed by Senators Richard Blumenthal (D-CT) and Josh Hawley (R-MO) requires AI companies to apply for licensing and also outlines measures to make it clear that Section 230's tech liability shield does not extend to generative AI (Blumenthal & Hawley 2023). Section 230(c)1 of the 1996 Telecommunications Act, which consists of 26 words—"No provider or user of an interactive computer service shall be treated as the publisher or speaker of

any information provided by another information content provider"—began as a catalyst for then–Vice President Al Gore's information superhighway but ended as a liability shield enabling Big Tech to behave arbitrarily on content moderation. It essentially treats social media as a platform rather than a publisher, even though the content the user sees is mediated by AI recommender systems. What this has meant in practice is that Big Tech, considered by law as a mere platform rather than a publisher, has not been liable for the harms its technology produces. The adverse effects of this legislation that has outgrown its usefulness are what the United States freely exports to the rest of the world (Lanier & Stanger 2024, Stanger 2024).

In addition to making it clear that companies will be held liable for the adverse effects of AI products, the Blumenthal-Hawley framework calls for a licensing regime administered by an independent oversight body. Companies would be required to register with the oversight authority in order to develop frontier AI models. The oversight authority would have the power to audit companies and prevent the licensing of those that prioritize profit over safety (Blumenthal & Hawley 2023).

Section 4.6 of the Biden administration's Executive Order on AI calls for input from various stakeholders on dual-use foundation models for which the model weights are widely available (White House 2023e). This solicitation process will be another contribution to an active, broader debate on the merits of open-sourcing foundation models. Dangers include enabling malicious use and spreading unresolved flaws, while benefits include increased external scrutiny, accelerating positive use cases, and reducing corporate control (Seger et al. 2023). Perhaps the most concerning abuse of open-source models is removing their guardrails for assistance with accessing dangerous pathogens; open-source AI models should be a prominent consideration in discussions of the dual-use nature of biological information (Gopal et al. 2023, Sandbrink 2023).

Beyond model weights, there is a broader question of how much information frontier AI developers should share about their models and development processes, and with whom. Transparency has some clear benefits; for example, access to cutting-edge models can support productive safety research (Bucknall & Trager 2023). One downside is sharing trade secrets with incautious or malicious actors, as well as nations conducting human rights abuses. Since companies will have less to gain from any single model when competitors can quickly replicate it, another issue is disincentivizing investments in large-scale models. These considerations will surely evolve over time as AI increases in capability (Seger et al. 2023). Developers will thus be required to navigate collective action problems and ensure robust cybersecurity practices when model information is intended to remain protected (Askell et al. 2019).

All regulation efforts to date have largely skirted the question of how to contain the risks of open-source models, which can proliferate virally in open societies via the current social media ecosystem. The free proliferation of open-source models, such as Meta's Llama and TII's Falcon, thus introduces an additional layer of complexity. In May 2023 on Stanford's campus, one of us could point our phone at a QR code on multiple public benches or billboards and download the weights needed to run Llama on a laptop. Improving AI safety will mean companies requiring a license to release open-source software built on LLMs and governments paying close attention to those companies deploying massive amounts of compute and stockpiling graphics processing units to scale existing models. Yet when these powerful models enter the wild, all bets are off.

One way to tackle this complex AI safety challenge is via digital supply chain monitoring. There needs to be a distinction between basic research and deployment. The monitoring should focus on strengthening safety requirements for bringing new frontier AI products to market. To that end, governments should (*a*) invest in digital supply chain monitoring, (*b*) invest in public evaluations of foundation models, and (*c*) incentivize research on guardrails for open-source models (Bommasani et al. 2023c, p. 1).

While industry self-regulation is a viable starting point, the challenge cannot be met without broad societal and multi-stakeholder involvement in a public process of deliberation (Landemore 2020). AI alignment assemblies using a consensus-generating deliberative digital tool, such as Pol.is, which was used in Taiwan's public debate about how to deal with ride-share culture in a world of licensed taxi drivers, hold promise. Since broad societal awareness coupled with government intervention will be required to ensure compliance, Pol.is is also being deployed in the exploration of AI alignment (Ovadya 2022, Tang & Landemore 2021). Put another way, "effective AI safety management requires transdisciplinary approaches and a shared language that allows involvement of all levels of society" (Dobbe 2022, p. 1). It will require both a multi-disciplinary and a multi-stakeholder approach.

According to European Commission President Ursula von der Leyen, Europe has placed its faith in a new international regime for managing frontier AI risk that is analogous to the UN Intergovernmental Panel on Climate Change. It would protect "against systemic societal risks and foster investments in safe and responsible AI systems, at the same time" (Von der Leyen 2023). In contrast, OpenAI's priorities for the governance of superintelligence are the establishment of an IAEA (International Atomic Energy Agency) equivalent for superintelligence and government investment in AI safety–related technical research (Fischer et al. 2021, Altman et al. 2023). Others suggest that a key policy action is to treat frontier AI as critical infrastructure (Smith et al. 2023). The need to do something to oversee the unfolding exponential growth in the power of foundation models is urgent (Anthropic 2023a), because research suggests that models will only become more capable as they get larger (Anthropic 2023b).

AI labs would also benefit from installing an internal audit function complemented by robust whistleblower protection principles. Rather than perpetual turnover in AI trust and safety personnel at the major companies, the pattern should instead be the development of an internal culture that is not single-mindedly devoted to the optimization of revenues. This is admittedly difficult to pull off in an economy that celebrates free markets and entrepreneurship (Stanger 2019).

A DUAL-TRACK STRATEGY

With unparalleled access to the algorithms, data, and graphics processing units that fuel advances in generative AI, the United States currently has a comparative advantage vis-à-vis China (Toner et al. 2023). At the same time, anti-China rhetoric is the one thing that unites both Democrats and Republicans in Washington, and it is matched by Chinese anti-American rhetoric. Unlike the American market-based and European rights-based approaches, the Chinese regulatory model is wholly state-driven—which means Communist Party–driven, as noted above. The United States and China each view the collective values of the other as a threat to their own. Not surprisingly, the Chinese and American versions of TikTok differ drastically. While American users consume videos that are predicted to engage them for as long as possible, the Chinese version (Douyin) limits use time so that teenagers can study and presents the Chinese user with a menu of videos meant to improve the mind and the user's sense of well-being and belonging (60 Minutes 2022, Schlott 2023).

Given China's party-controlled surveillance state, if the United States were to win an arms race with China on Chinese terms, that victory would destroy what Americans endeavor to defend. Yet, opportunities are present for constructive engagement with China on AI regulation. Unrestrained great power rivalries are at the heart of extreme risk dynamics. Building capable global institutions is perhaps the only way to defend US national security while containing the myriad potential negative consequences of unrestrained competition that will benefit no one. "Superiority is not synonymous with security" (Danzig 2018). Any strategy for cooperation with China,

therefore, must incorporate how AI is likely to affect the national power of individual states as well as features of the international system (Bajraktari et al. 2023). For example, there is evidence that AI's destabilizing effects are "likely to be more strongly de-democratizing in emerging and peripheral economies," where states simply lack the capacity to protect their populations from deindustrialization (Boix 2022, p. 1).

A dual-track strategy for AI governance that strengthens ties between open societies, while building institutions and norms for cooperation on areas of mutual concern in the broader global community, is therefore both realistic and prudent. Requiring China to adhere to Western values or the range of different AI principles that have been promulgated as a necessary condition for communication between the United States and China serves the interests of neither country. The legitimating principles of the American and Chinese political systems differ dramatically, so the range of possible options for each is constrained quite differently (Stanger 2020, 2021). Free-world policymakers should explore China's willingness to participate in binding international negotiations that in turn can serve, much as arms control negotiations did during the Cold War, as confidence-building measures minimizing the potential for miscommunication that leads to catastrophe (Horowitz & Scharre 2021, Horowitz et al. 2022).

For example, global governance of AI might be pursued through a new global agency—what some have referred to as the functional equivalent of an IAEA for AI (Maas & Villalobos 2023, Trager et al. 2023), as mentioned above. In such a body, members might pledge no first use of cyber weapons against domestic critical infrastructure in exchange for respect of the right to self-determination within a UN-recognized state's borders. Members could also pledge no first use of drone armies. One could also envision such an institution monitoring the global flows of GPUs to ensure they do not fall into dangerous hands, the latter to be collectively defined by the signatories to the new international regime. Some see the Financial Action Task Force as a good model for cloud compute know-your-customer provisions (Egan & Heim 2023, Fist et al. 2023). The G7 and OECD's Global Partnership on AI (**https://oecd.ai/en/gpai**) are existing multi-stakeholder global institutions that might also serve to bolster guardrails for governing tumultuous innovation. Finally, some have argued for a US–European CERN for AI, analogous to the European Organization for Nuclear Research, that might increase publicly funded cutting-edge frontier AI research (Hoos & Irgens 2023).

The issue of who has access to the massive amount of compute needed to fuel advances in deep learning and generative AI is not likely to go away any time soon (Vipra & Myers West 2023). A dual-track approach increases the number of points of contact between rivals while simultaneously strengthening allied cooperation, which has historically proven to be useful for avoiding existential-risk-magnitude forms of conflict sparked by simple miscommunication or malfunctioning hardware or software (Jervis 1997).

Adopting a dual-track approach to AI and machine learning governance is also a good way to address the dueling perspectives on AI and the US–China relationship, what Ding (2022) calls the technonationalism versus technoglobalism divide. Technonationalism emphasizes interstate competition, whereas technoglobalism emphasizes technological assets. By focusing on the transnational networks of firms and individuals that technoglobalism highlights, we can discern a clearer picture of where the United States and China might find both discord and common ground (Ding 2022). There are areas in which the free world and China can reach agreements because of their complex interdependence. But there are also areas where we cannot, due to serious national security issues. Knowing the difference is a key insight for building new governance institutions that can protect both freedom's and humanity's interests.

The technoglobalist perspective suggests that there may be substantive governance gains from refocusing our attention from the models to their inputs. Frontier models are built on troves of

data. Developers can only develop next-generation models insofar as they have a stable, reliable, and high-quality data supply chain, a supply chain which, since it exists on the internet (a global technology), is not confined to national borders. Perhaps, then, it might be helpful to think about these frontier models, the models that rely on vast data supply chains, as problems in ethical and sustainable supply chain management. There is already a significant body of literature on the ways national and international organizations can enforce ethical, sustainable, and prolabor sourcing standards (McKean 2022), and although we do not yet have a clean solution to those problems, that work is at least a good starting point for thinking through the unique challenges that AI supply chains pose, especially when open-society allies effectively control the supply chain for GPU chips (Belfield 2022, Miller 2022).

It follows that international and transnational regimes may be able to play an important role in ensuring that frontier AI systems benefit humanity rather than destroying it. "Four potential institutional models that have precedents in existing organizations to consider are: 1) a Commission on Frontier AI that facilitates expert consensus on opportunities and risks from advanced AI, 2) an Advanced AI Governance Organization that sets international standards to manage global threats from advanced models, supports their implementation, and possibly monitors compliance with a future governance regime, 3) a Frontier AI Collaborative that promotes access to cutting-edge AI, and 4) an AI Safety Project that brings together leading researchers and engineers to further AI safety research" (Ho et al. 2023, p. 1).

A dual-track strategy (see **Table 1**) that builds new global governance regimes while strengthening cooperation and coordination among open societies can reduce the likelihood of catastrophe while rendering liberal democracy itself sustainable. It does so by creating multiple intervention possibilities for containing the unexpected—a likely outcome when open-source systems are fully unleashed in the wild, where a few malevolent actors can wreak havoc. Cooperation between adversaries also reduces the risk of World War III with new and "improved" weapons (some involving AI, like drones) that might cause far more destruction than World War II, especially if nuclear powers were to utilize their arsenals (Boulanin et al. 2020).

Table 1 Approaches to artificial intelligence (AI) regulation

	United States	European Union	China
Core AI regulation approach	Market-driven	Rights-driven	State-driven
Governing institutions	Bicameral representative democracy	Multinational parliamentary representative democracy	Single-party autocracy
User rights and privacy policy	Limited regulation, history of both public and private sector user-data collection and surveillance	Increasingly regulated (GDPR, Digital Markets Act, etc.), limits on public and private surveillance	Systematic data aggregation and surveillance, limited user access to international internet activity
Response to AI risks	Democracy/rights promotion, open internet	Democracy/rights promotion, open internet	Strengthen party control
Policy action to date	AI Bill of Rights, NIST AI Risk Management Framework, CHIPS and Science Act, Executive Order on AI	AI Act, GDPR, DMA, DSA	Regulations on recommendation algorithms, deep synthesis, and generative AI; party oversight of publicly held LLMs
We propose	Dual-track approach	Dual-track approach	Self-determination and global AI governance initiatives

Abbreviations: CHIPS, Creating Helpful Incentives to Produce Semiconductors; DMA, Digital Markets Act; DSA, Digital Services Act; GDPR, General Data Protection Regulation; LLM, large language model; NIST, National Institute of Standards and Technology.

CONCLUSION: AI SAFETY IS AI GOVERNANCE

The computer science community sounded the alarm on AI existential risk and the need for government intervention with the Future of Life Institute's March 2023 pause petition, which called on all AI labs to cease training AI systems more powerful than GPT4 for 6 months (Future of Life Inst. 2023). The pause petition was followed 2 months later by the Center for AI Safety's one-sentence petition: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." It was signed by many players in the general field, capturing the world's attention (Cent. AI Safety 2023). In early November 2023, representatives from 28 countries, including the United States and China, convened at the UK AI Safety Summit and agreed on the Bletchley Declaration, which stated, "There is potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of these AI models" (AI Safety Summit 2023).

The call for taking a breath and assessing the risks was understandable. As we have seen, a daunting number of interdependent, complex, adaptive systems constitute the AI landscape. Sustaining democracy with this landscape as a backdrop and resisting the techno-authoritarian alternative that delivers greater efficiency are going to require a societal response. One way of summarizing what we have learned from the previous pages is to capture it with a typology of instruments for building out systems that uphold human dignity rather than trampling upon it. Those instruments are (*a*) rules, such as new regulations and laws; (*b*) new institutions, which can reinforce, amplify, and harmonize the rules; (*c*) money, whether to fund research, invest in AI safety, or increase funding for education of a technologically competent citizenry; and (*d*) human capital, the people who can imagine solutions to and mitigate risks from the challenges of further technological innovation (Matthews 2023).

The reader may have noted that we have not directly taken up arguments for rendering generative AI more transparent or explainable. This is because both transparency and explainability are devilishly challenging concepts, distinct from governance. Although explainability may be a useful tool in the design and implementation of AI governance mechanisms, explainable AI (XAI) systems are not necessarily more governable than non-XAI systems (Danks 2022). The right to an "explanation," therefore, could be answered with more powerful AI, but this would simply amount to ceding still more human agency to machines. Human beings, after all, excel at *ex post* explanations—another word for individuation or developing one's own narrative—just as these machines could be trained by humans to do. "If we put too much emphasis on explainability," says Edward Lee, Distinguished Professor in Electrical Engineering and Computer Sciences at the University of California, Berkeley, "we will undermine human contestability" (Lee 2023). If we want to reinforce democracy rather than undermine it, it may be better to retain the right to appeal to a human rather than relying on machines inventing explanations for the inexplicable (Rees 2022).

Similarly, the excessive focus on technical misalignment risks can blind us to the ways in which unregulated frontier AI models are likely to diminish humans rather than expanding them in the immediate future. Existential risks are real, but the immediate clear and present danger is democracy's demise. From September 2023 through November 2024, i.e., in the age of largely unregulated frontier AI, 80% of democratic countries (representing 67% of all people who live in a democracy) will vote in a federal or supranational (e.g., European Parliament) election. The threat to democracy of automated disinformation in a global social media ecosystem increasingly controlled by one man, Mark Zuckerberg, is grave. The most urgent threat to human flourishing is the one staring all of us in the face, right here and now.

Drawing attention to the ways in which AI systems are used to exercise power ultimately demonstrates the inadequacy of AI ethics and principles approaches alone. "When new and intensified power relations develop, we must attend not only to what power is used for, but also to how and by whom it is used" (Lazar 2022, p. 1). Limiting and reversing the human costs of AI technologies "may need to rely on regulation and policies to redirect AI research. Attempts to contain them just by promoting competition may be insufficient" (Acemoğlu 2021, p. 1). In short, the limits of market solutions have been reached (Acemoğlu 2022).

The interesting question is whether the spark of creativity or intuition, the product of human memory, will one day be generated by machine computation. One might privilege the human brain, but we know that its complex structure is powered by both chemistry and electricity. So, too, are machines built from chips. We will learn more about human consciousness by studying generative AI, and our understanding of both will likely develop in tandem.

An explanation, after all, is a short algorithm with steps upon which society has agreed. Might it not be the case, then, that our human brains are already performing algorithms, as the computational theories of mind have already suggested (Rescorla 2020)? The mistakes made by GPT2, GPT3 (and GPT3.5), and GPT4 remind Berkeley professor Edward Lee (2023) of the mistakes of a 4-year-old, a 15-year-old, and a Berkeley graduate student—and this stellar progression in capabilities took place in less than 5 years, with a significant share of progress coming from training with orders of magnitude more computational power. We are left to wonder whether the gap between human ingenuity and AI computation will vanish with time or whether human embodied intelligence will continue to have the upper hand. Given the possibility that future change could occur still more rapidly, it is wise to prepare for more surprises (Erdil & Besiroglu 2023, Davidson 2023).

We hope this state-of-the-field review inspires other researchers to explore the complex, adaptive *terra incognita* of global AI governance and to interact with our arguments. We encourage you to tell us why we are wrong. Promising research questions include:

- How do public institutions govern systems under conditions of deep asymmetries of power (both within the tech sector and across tech and government) in terms of the design of code (informational power) and the control of it (market power)?
- How do they do so in ways that allow accountability—which requires transparency—in proprietary systems that are opaque by design?
- How do they do so when the speed of innovation is much faster than the organizational changes in governing institutions?
- How do they do so when the technology may be deployed in ways that systematically and intentionally undermine these same governing institutions (threats to democracy)?
- How do they do so when technology poses a potential deep existential risk?

In the end, sustaining democratic governance is itself an AI safety question, which means that AI safety is more than a technical challenge. If the disempowered come to see AI as a tool for their own subordination, many will feel justified in exploiting AI to make the powerful aware of their existence. The desire to be seen and valued is a human need that machines will struggle to compute away. Without citizens who see themselves reflected in their governments, AI safety, broadly speaking, will be an elusive goal.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank the GETTING-Plurality Research Network, the study group on Complexity and International Politics (Princeton/New America/Arizona State University), Danielle Allen, Barbara Grosz, Zoe Hitzig, Eric Horvitz, Jaron Lanier, Edward Lee, Colin Megill, James Landay, James Manyika, Margaret Levi, Melanie Mitchell, Cris Moore, Charlotte Siegmann, Moshe Vardi, Glen Weyl, David Wolpert, and an anonymous reviewer for helpful comments and conversations.

LITERATURE CITED

60 Minutes. 2022. TikTok in China versus the United States. *60 Minutes*, Nov. 8. **https://www.youtube.com/watch?v=0j0xzuh-6rY**

Acemoğlu D. 2021. *Harms of AI*. NBER Work. Pap. 29247

Acemoğlu D. 2022. Harms of AI. See Bullock et al. 2022, **https://doi.org/10.1093/oxfordhb/9780197579329.013.65**

AI Safety Summit. 2023. The Bletchley Declaration by countries attending the AI Safety Summit, 1–2 November 2023. **https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023**

Allen D, Hubbard S, Lim W, Stanger A, Wagman S, Zalesne K. 2024. *A roadmap for governing AI: technology governance and power sharing liberalism*. Work. Pap., Ash Cent. Democr. Gov. Innov., Harvard Kennedy School, Cambridge, MA. **https://www.youtube.com/watch?v=0j0xzuh-6rYhttps://ash.harvard.edu/publications/roadmap-governing-ai-technology-governance-and-power-sharing-liberalism**

Altman S, Brockman G, Sutskever I. 2023. Governance of superintelligence. *OpenAI*, May 22. **https://openai.com/blog/governance-of-superintelligence**

Amodei D. 2023. Written testimony of Dario Amodei, PhD, co-founder and CEO, Anthropic: for a hearing on "Oversight of A.I.: Principles for Regulation" before the Judiciary Committee Subcommittee on Privacy, Technology, and the Law. US Senate, July 25. **https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf**

Anderljung M, Barnhart J, Kolinek A, Leung J. 2023. Frontier AI regulation: managing emerging risks to public safety. arXiv:2307.03718 [cs.CY]

Anderljung M, Hazell J. 2023. Protecting society from AI misuse: When are restrictions on capabilities warranted? arXiv:2303.09377 [cs.AI]

Anthropic 2023a. Frontier model security. *Anthropic*, July 25. **https://www.anthropic.com/news/frontier-model-security**

Anthropic 2023b. Frontier threats: red teaming for AI safety. *Anthropic*, July 26. **https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety**

Arsenault A, Kreps S. 2022. AI and international politics. See Bullock et al. 2022, **https://doi.org/10.1093/oxfordhb/9780197579329.013.49**

Askell A, Brundage M, Hadfield G. 2019. The role of cooperation in responsible AI development. arXiv:1907.04534 [cs.CY]

Bajraktari Y, Gable M, Ponmakha A. 2023. *Generative AI: the future of innovation power*. Rep. Spec. Compet. Stud. Proj., Arlington, VA. **https://www.scsp.ai/wp-content/uploads/2023/09/GenAI-web.pdf**

Barrett AM, Hendrycks D, Newman J, Nonnecke B. 2023. Actionable guidance for high-consequence AI risk management: towards standards addressing AI catastrophic risks. arXiv:2206.08966 [cs.CY]

Belfield H. 2022. Compute and antitrust. *Verfassungsblog: On Matters Constitutional*, Aug. 19. **https://verfassungsblog.de/compute-and-antitrust/**

Blumenthal R, Hawley J. 2023. Bipartisan framework for U.S. AI Act. Sep. 7. *Richard Blumenthal, U.S. Senator for Connecticut.* **https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf**

Boix C. 2022. AI and the economic and informational foundations of democracy. See Bullock et al. 2022, **https://doi.org/10.1093/oxfordhb/9780197579329.013.64**

Bommasani R, Hashimoto T, Ho D, Schaake M, Liang P. 2023b. *Towards compromise: a concrete two-tier proposal for foundation models in the EU AI Act*. Work. Pap., Cent. Res. Found. Models, Stanford Univ., Stanford, CA. **https://crfm.stanford.edu/2023/12/01/ai-act-compromise.html**

Bommasani R, Kapoor S, Zhang D, Narayanan A, Liang P. 2023c. *June 12 letter to the Department of Commerce*. Cent. Res. Found. Models, Stanford Univ., Stanford, CA. **https://hai.stanford.edu/sites/default/files/2023-06/Reponse-to-NTIAs-.pdf**

Bommasani R, Klyman K, Zhang D, Liang P. 2023a. *Do foundation model providers comply with the draft EU AI Act?* Work. Pap., Cent. Res. Found. Models, Stanford Univ., Stanford, CA

Boulanin V, Saalman L, Topychkanov P, Su F, Peldán C. 2020. *Artificial Intelligence, Strategic Stability and Nuclear Risk*. Stockholm: Stockholm Int. Peace Res. Inst. **https://www.sipri.org/publications/2020/policy-reports/artificial-intelligence-strategic-stability-and-nuclear-risk**

Bradford A. 2020. *The Brussels Effect: How the European Union Rules the World*. New York: Oxford Univ. Press

Bradford A. 2023a. *Digital Empires: The Global Battle to Regulate Technology*. New York: Oxford Univ. Press

Bradford A. 2023b. The race to regulate artificial intelligence. *Foreign Aff.*, June 27. **https://www.foreignaffairs.com/united-states/race-regulate-artificial-intelligence**

Bremmer I, Suleyman M. 2023. The AI power paradox. *Foreign Aff.*, Aug. 16. **https://www.foreignaffairs.com/world/artificial-intelligence-power-paradox**

Bucknall B, Trager R. 2023. *Structured access for third-party research on frontier AI models: investigating researchers' model access requirements*. Rep., Cent. Gov. AI, Oxford Univ., Oxford, UK. **https://cdn.governance.ai/Structured_Access_for_Third-Party_Research.pdf**

Bullock JB, Chen Y-C, Himmelreich J, Hudson VM, Korinek A, et al., eds. 2022. *The Oxford Handbook of AI Governance*. New York: Oxford Univ. Press

Casper S, Davies X, Shi C, Gilbert TK, Scheurer J, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv:2307.15217 [cs.AI]

Cent. AI Safety. 2023. Statement on AI risk. *Cent. AI Safety.* **https://www.safe.ai/statement-on-ai-risk**

Clement-Jones L. 2023. The Westminster Parliament's impact on UK AI strategy. In *Missing Links in AI Governance*, ed. B Prud'homme, C Régis, G Farnadi, pp. 191–209. Paris: UNESCO

Dafoe A. 2018. *AI governance: a research agenda*. Rep. Future Humanit. Inst., Univ. Oxford, Oxford, UK. **https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf**

Dafoe A. 2022. AI governance: overview and theoretical lenses. See Bullock et al. 2022, **https://doi.org/10.1093/oxfordhb/9780197579329.013.2**

Danks D. 2022. Governance via explainability. See Bullock et al. 2022, **https://doi.org/10.1093/oxfordhb/9780197579329.013.11**

Danzig R. 2018. *Technology roulette: managing loss of control as many militaries pursue technological superiority*. Rep. Cent. New Am. Secur., Washington, DC. **https://www.cnas.org/publications/reports/technology-roulette**

Davidson T. 2023. What a compute-centric framework says about takeoff speeds. *Lesswrong Online Forum*, Jan. 22. **https://www.lesswrong.com/posts/Gc9FGtdXhK9sCSEYu/what-a-compute-centric-framework-says-about-ai-takeoff**

Ding J. 2022. Dueling perspectives in AI and U.S.–China relations: technonationalism vs. technoglobalism. See Bullock et al. 2022, **https://doi.org/10.1093/oxfordhb/9780197579329.013.53**

Dobbe R. 2022. System safety and artificial intelligence. See Bullock et al. 2022, **https://doi.org/10.1093/oxfordhb/9780197579329.013.67**

Egan J, Heim L. 2023. Oversight for frontier AI through a know-your-customer scheme for compute providers. arXiv:2310.13625 [cs.CY]

Erdil E, Besiroglu T. 2023. Explosive growth from AI automation: a review of the arguments. arXiv:2309.11690 [econ.GN]

Espinoza J, Criddle C, Liu Q. 2023. The global race to set the rules for AI. *Financ. Times*, Sep. 13. **https://www.ft.com/content/59b9ef36-771f-4f91-89d1-ef89f4a2ec4e**

European Parliament. 2024. *The principle of subsidiarity*. Fact Sheet Eur. Union. **https://www.europarl.europa.eu/erpl-app-public/factsheets/pdf/en/FTU_1.2.2.pdf**

Fischer S-C, Leung J, Anderljung M, O'Keefe C, Torges S, et al. 2021. *AI policy levers: a review of the U.S. government's tools to shape AI research, development, and deployment*. Rep. Cent. Gov. AI, Future Humanit. Inst.,

Univ. Oxford, Oxford, UK. **https://www.fhi.ox.ac.uk/wp-content/uploads/2021/03/AI-Policy-Levers-A-Review-of-the-U.S.-Governments-tools-to-shape-AI-research-development-and-deployment—-Fischer-et-al.pdf**

Fist T, Heim L, Schneider J. 2023. Chinese firms are evading chip controls. *Foreign Policy*, June 21. **https://foreignpolicy.com/2023/06/21/china-united-states-semiconductor-chips-sanctions-evasion/**

Future of Life Inst. 2023. Pause giant AI experiments: an open letter. *Future of Life*, Mar. 22. **https://futureoflife.org/open-letter/pause-giant-ai-experiments/**

Ganguli D, Hernandez D, Lovitt L, DasSarma N, Henighan T, et al. 2022. Predictability and surprise in large generative models. In *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1747–64. New York: Assoc. Comput. Mach. **https://dl.acm.org/doi/pdf/10.1145/3531146.3533229**

Gopal A, Helm-Burger N, Justen L, Soice EH, Tzeng T, et al. 2023. Will releasing the weights of future large language models grant widespread access to pandemic agents? arXiv:2310.18233 [cs.AI]

Graham L, Warren E. 2023. When it comes to Big Tech, enough is enough. *N. Y. Times*, July 27. **https://www.nytimes.com/2023/07/27/opinion/lindsey-graham-elizabeth-warren-big-tech-regulation.html**

Gray ML, Suri S. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt

Guszcza J, Danks D, Fox CR, Hammond KJ, Ho DE, et al. 2022. *Hybrid intelligence: a paradigm for more responsible practice*. Soc. Sci. Res. Netw. **http://dx.doi.org/10.2139/ssrn.4301478**

gwern. 2023. Douglas Hofstadter changes his mind on deep learning & AI risk. *Lesswrong Online Forum*, July 2. **https://www.lesswrong.com/posts/kAmgdEjq2eYQkB5PP/douglas-hofstadter-changes-his-mind-on-deep-learning-and-ai**

Harari Y. 2017. Reboot for the AI revolution. *Nature* 550:324–27

Hawley J. 2021. *The Tyranny of Big Tech*. Washington, DC: Regnery

Heidegger M. 1977. *The Question Concerning Technology, and Other Essays*., transl. W Lovitt. New York: Garland

Ho L, Barnhart J, Trager R, Bengio Y, Brundage M, et al. 2023. International institutions for advanced AI. arXiv:2307.04699 [cs.CY]

Hoffman M. 2023. The EU AI Act: a primer. *CSET Blog*, Sep. 26, Cent. Secur. Emerg. Technol., Georgetown Univ., Washington, DC. **https://cset.georgetown.edu/article/the-eu-ai-act-a-primer/**

Hoos H, Irgens M. 2023. *"AI made in Europe"—boost it or lose it*. CLAIRE Statement on Future of AI in Europe, June 26. Confed. Lab. Artif. Intell. Eur., The Hague, Neth. **https://claire-ai.org/wp-content/uploads/2023/06/CLAIRE-Statement-on-Future-of-AI-in-Europe-2023-1.pdf**

Horowitz M, Pindyck S, Mahoney C. 2022. AI, the international balance of power, and national security strategy. See Bullock et al. 2022, **https://doi.org/10.1093/oxfordhb/9780197579329.013.55**

Horowitz M, Scharre P. 2021. *AI and international stability: risks and confidence-building measures*. Rep. Cent. New Am. Secur., Jan. 12. **https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures**

Jervis R. 1997. *System Effects: Complexity in Political and Social Life*. Princeton, NJ: Princeton Univ. Press

Kauffman S. 2008. *Reinventing the Sacred: A New View of Science, Reason, and Religion*. New York: Basic Books

Khan LM. 2017. Amazon's antitrust paradox. *Yale Law J.* 126(3):710–805

Knight W. 2023. OpenAI's CEO says the age of giant AI models is already over. *WIRED*, Apr. 17. **https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/**

Kokas A. 2023. *Trafficking Data: How China Is Winning the Battle for Digital Sovereignty*. New York: Oxford Univ. Press

Kurzweil R. 2006. *The Singularity Is Near: When Humans Transcend Biology*. New York: Penguin

Landemore H. 2020. *Open Democracy: Reinventing Popular Rule for the Twenty-First Century*. Princeton, NJ: Princeton Univ. Press

Lanier J, Stanger A. 2024. The one internet hack that could save everything. *Wired*, Feb. 13. **https://www.wired.com/story/the-one-internet-hack-that-could-save-everything-section-230/**

Lanier J, Weyl EG. 2018. A blueprint for a better digital society. *Harvard Bus. Rev.*, Sep. 26. **https://hbr.org/2018/09/a-blueprint-for-a-better-digital-society**

Larsen B. 2022. The geopolitics of AI and the rise of digital sovereignty. *Brookings*, Dec. 8. **https://www.brookings.edu/articles/the-geopolitics-of-ai-and-the-rise-of-digital-sovereignty/**

Lazar S. 2022. Power and AI: nature and justification. See Bullock et al. 2022, **https://doi.org/10.1093/oxfordhb/9780197579329.013.12**

Lee EA. 2023. *Edward A. Lee: Deep neural networks, explanations, and rationality*. Lecture presented in 2nd Digital Humanism Summer School 2023, Sep. 6. **https://www.youtube.com/watch?v=yva7kPCQ2lc**

Lessig L. 2006. *Code: Version 2.0*. New York: Basic Books

Lipton E. 2023. A.I. brings the robot wingman to aerial combat. *N. Y. Times*, Aug. 27. **https://www.nytimes.com/2023/08/27/us/politics/ai-air-force.html**

Maas MM, Villalobos JJ. 2023. *International AI institutions: a literature review of models, examples, and proposals*. Soc. Sci. Res. Netw. **https://doi.org/10.2139/ssrn.4579773**

Maslej N, Fattorini L, Brynjolfsson E, Etchemendy J, Ligett K, et al. 2023. *The AI Index 2023 annual report*. Rep. Inst. Hum.-Cent. Artif. Intell., Stanford Univ., Stanford, CA. **https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf**

Matthews D. 2023. The AI rules that US policymakers are considering, explained. *Vox*, Aug. 1. **https://www.vox.com/future-perfect/23775650/ai-regulation-openai-gpt-anthropic-midjourney-stable**

McKean BL. 2022. *Disorienting Neoliberalism: Global Justice and the Outer Limit of Freedom*. New York: Oxford Univ. Press

Microsoft. 2023. Microsoft, Anthropic, Google, and OpenAI launch Frontier Model Forum. *Microsoft On the Issues Blog*, Jul. 26. **https://blogs.microsoft.com/on-the-issues/2023/07/26/anthropic-google-microsoft-openai-launch-frontier-model-forum/**

Miller C. 2022. *Chip Wars: The Fight for the World's Most Critical Technology*. New York: Scribner

Mitchell M 2020. *Artificial Intelligence: A Guide for Thinking Humans*. New York: Picador

Mitchell M. 2023. Can large language models reason? *AI: A Guide for Thinking Humans Substack*, Sep. 10. **https://aiguide.substack.com/p/can-large-language-models-reason**

Mullainathan S, Obermeyer Z. 2021. On the inequity of predicting A while hoping for B. *AEA Pap. Proc.* 111:37–42

Noble S. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press

O'Neil C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown

Ord T. 2022. *Lessons from the development of the atomic bomb*. Rep. Cent. Gov. AI, Oxford Univ., Oxford, UK. **https://cdn.governance.ai/Ord_lessons_atomic_bomb_2022.pdf**

Ovadya A. 2022. *Bridging-based ranking*. Rep. Belfer Cent. Sci. Int. Aff., Harvard Kennedy Sch., Cambridge, MA. **https://www.belfercenter.org/sites/default/files/files/publication/TAPP-Aviv_BridgingBasedRanking_FINAL_220518_0.pdf**

Park PS, Goldstein S, O'Gara A, Chen M, Hendrycks D. 2023. AI deception: a survey of examples, risks, and potential solutions. arXiv:2308.14752 [cs.CY]

Patel D, Ahmad A. 2023. Google "we have no moat, and neither does OpenAI." *SemiAnalysis*, May 4. **https://www.semianalysis.com/p/google-we-have-no-moat-and-neither**

Pearson M, Rithmire M, Tsai K 2022. China's party-state capitalism and international backlash: from interdependence to insecurity. *Int. Secur.* 47(2):135–76

Rees T. 2022. Non-human words: on GPT-3 as a philosophical laboratory. *Daedalus* 151:168–82

Rescorla M. 2020. The computational theory of mind. In *Stanford Encyclopedia of Philosophy Archive*, ed. EN Zalta. **https://plato.stanford.edu/archives/fall2020/entries/computational-mind/**

Russell S. 2022. Banning lethal autonomous weapons: an education. *Issues Sci. Technol.* Spring:60–65. **https://issues.org/wp-content/uploads/2022/04/60-65-Russell-Lethal-Autonomous-Weapons-Spring-2022.pdf**

Sandbrink JB. 2023. Artificial intelligence and biological misuse: differentiating risks of language models and biological design tools. arXiv:2306.13952 [cs.CY]

Schaeffer R, Miranda B, Koyejo S. 2023. Are emergent abilities of large language models a mirage? arXiv:2304.15004 [cs.AI]

Schlott L. 2023. China is hurting our kids with TikTok but protecting its own youth with Douyin. *N. Y. Post*, Feb. 26. **https://nypost.com/2023/02/25/china-is-hurting-us-kids-with-tiktok-but-protecting-its-own/**

Seger E, Dreksler N, Moulange R, Dardaman E, Schuett J, et al. 2023. *Open-sourcing highly capable foundation models.* Rep. Cent. Gov. AI, Oxford Univ., Oxford, UK. **https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf**

Sheehan M. 2023. *Reverse engineering Chinese AI governance: China's AI regulations and how they get made*. Rep. Carnegie Endow. Int. Peace, Washington, DC. **https://carnegieendowment.org/files/202307-Sheehan_Chinese%20AI%20gov.pdf**

Shevlane T, Farquhar S, Garfinkel B, Phuong M, Whittlestone J, et al. 2023. Model evaluation for extreme risks. arXiv:2305.15324 [cs.AI]

Smith G, Kessler S, Alstott J, Mitre J. 2023. *Industry and Government Collaboration on Security Guardrails for AI Systems: Summary of the AI Safety and Security Workshops*. Santa Monica, CA: RAND

Soice EH, Rocha R, Cordova K, Specter M, Esvelt KM. 2023. Can large language models democratize access to dual-use biotechnology? arXiv:2306.03809 [cs.CY]

Stanger A. 2019. *Whistleblowers: Honesty in America from Washington to Trump*. New Haven, CT: Yale Univ. Press

Stanger A. 2020. Consumers versus citizens in democracy's public sphere. *Commun. ACM* 63(7):29–31

Stanger A. 2021. *Digital humanism and democracy in geopolitical context*. Digital Humanism Sem. Ser., TU Wien, June 15. **https://www.youtube.com/watch?v=ELeR1ViNiUQ**

Stanger A. 2024. The First Amendment meets Big Tech. *Daedalus*. In press

Stiglitz JE. 2002. *Globalization and Its Discontents*. New York: W. W. Norton

Strittmatter K. 2020. *We Have Been Harmonized: Life in China's Surveillance State*. New York: Custom House

Susskind RE, Susskind D. 2022. *The Future of the Professions: How Technology Will Transform the Work of Human Experts*. Oxford, UK: Oxford Univ. Press

Tang A, Landemore H. 2021. *Taiwan's digital democracy, collaborative civic technologies, and beneficial information flows.* Rep. Cent. Gov. AI, Oxford Univ., Oxford, UK, Mar. 9. **https://www.governance.ai/post/audrey-tang-and-helene-landemore-on-taiwans-digital-democracy-collaborative-civic-technologies-and-beneficial-information-flows**

Teachout Z. 2020. *Break 'Em Up: Recovering Our Freedom from Big Ag, Big Tech, and Big Money*. New York: All Points Books

Toner H, Xiao J, Ding J. 2023. The illusion of China's AI prowess. *Foreign Aff.*, June 2. **https://www.foreignaffairs.com/china/illusion-chinas-ai-prowess-regulation**

Touvron H, Martin L, Stone K, Albert P, Almahairi A, et al. 2023. Llama 2: open foundation and fine-tuned chat models. arXiv:2307.09288 [cs.CL]

Trager R, Harack B, Reuel A, Carnegie A, Heim L, et al. 2023. International governance of civilian AI: a jurisdictional certification approach. arXiv:2308.15514 [cs.AI]

Viehoff J. 2022. Beyond justice: artificial intelligence and the value of community. See Bullock et al. 2022, **https://doi.org/10.1093/oxfordhb/9780197579329.013.71**

Vipra J, Myers West S. 2023. *Computational Power and AI*. AI Now Inst. **https://ainowinstitute.org/wp-content/uploads/2023/09/AI-Now_Computational-Power-an-AI.pdf**

Von der Leyen U. 2023. Statement by President von der Leyen at Session III of the G20, "One Future." Eur. Comm., Sep. 10, New Delhi. **https://ec.europa.eu/commission/presscorner/detail/en/statement_23_4424**

von Neumann J. 1963. *Collected Works*, Vol. VI, ed. A Taub, pp. 504–19. Oxford, UK: Pergamon

Vredenburgh K. 2022. Fairness. See Bullock et al. 2022, **https://doi.org/10.1093/oxfordhb/9780197579329.013.8**

Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, et al. 2022. Emergent abilities of large language models. arXiv:2206.07682 [cs.CL]

Weissinger L. 2022. AI, complexity, and regulation. See Bullock et al. 2022, **https://doi.org/10.1093/oxfordhb/9780197579329.013.66**

White House. 2023a. *Biden-Harris administration launches artificial intelligence cyber challenge to protect America's critical software*. Statement, Aug. 9, Briefing Room, White House, Washington, DC. **https://www.whitehouse.gov/briefing-room/statements-releases/2023/08/09/biden-harris-administration-launches-artificial-intelligence-cyber-challenge-to-protect-americas-critical-software/**

White House. 2023b. *Blueprint for an AI Bill of Rights.* White Pap., Off. Sci. Technol. Policy, White House, Washington, DC. **https://www.whitehouse.gov/ostp/ai-bill-of-rights/**

White House. 2023c. *President Biden issues executive order on safe, secure, and trustworthy artificial intelligence*. Fact Sheet, Oct. 30, Briefing Room, White House, Washington, DC. **https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/**

White House. 2023d. *Biden-Harris administration secures voluntary commitments from leading artificial intelligence companies to manage the risks posed by AI*. Fact Sheet, July 21, Briefing Room, White House, Washington, DC. **https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/**

White House. 2023e. *Executive Order on the safe, secure, and trustworthy development of artificial intelligence*. Exec. Order, Oct. 30, Briefing Room, White House, Washington, DC. **https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/**

Zuboff S. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs