AI & SOCIETY
https://doi.org/10.1007/s00146-024-02050-7

OPEN FORUM

International governance of advancing artificial intelligence

Nicholas Emery-Xu[2] · Richard Jordan[1] · Robert Trager[3]

Received: 27 March 2024 / Accepted: 6 August 2024
© The Author(s) 2024

Abstract

New technologies with military applications may demand new modes of governance. In this article, we develop a taxonomy of technology governance forms, outline their strengths, and red-team their weaknesses. In particular, we consider the challenges and opportunities posed by advancing artificial intelligence, which is likely to have substantial dual-use properties. We conclude that subnational governance, though prevalent and mitigating some risks, is insufficient when the individual rewards from societally harmful actions outweigh normative sanctions, as is likely to be the case with AI. Nationally enforced standards are promising ways to govern AI deployment, but they are less viable in the "race-to-the-bottom" environments that are becoming common. When it comes to powerful technologies with military implications, there is only one multilateral option with a strong historical precedent: a non-proliferation plus norms-of-use regime, which we call NPT+. We believe that a non-proliferation regime may, therefore, be the necessary foundation for AI governance. However, AI may exhibit characteristics that would make a non-proliferation regime less effective than it has proven for nuclear weapons. As an alternative, verification-backed restrictions on AI development and use would address more risks, but they face challenges in the case of advanced AI, and we show how these challenges may not have technical solutions. Perhaps more importantly, we show that there is no clear example of major powers restricting the development of a powerful military technology when that technology lacks a ready substitute. We, therefore, turn to a final alternative, International Monopoly, which was the preferred solution of many scholars and policymakers in the early nuclear era. It should be considered again for governing AI: a monopoly would require less-invasive monitoring, though at the possible cost of eroding national sovereignty. Ultimately, we conclude that it is too soon to tell whether a non-proliferation regime, a verification-based regime, or an International Monopoly is most feasible for governing AI. Nonetheless, a variety of policies would yield a high return across all three scenarios, and we conclude by identifying some of these steps that could be taken today.

Keywords AI governance · Non-proliferation · Risk · Institutions

Researchers at Collaborations Pharmaceuticals, a small drug company in Raleigh, NC, used artificial intelligence (AI) techniques to search for toxic molecules. After a few hours, they found 40,000 potential toxins. Some were known toxins, like the nerve agent VX, the most toxic chemical yet discovered, but many were predicted to be orders of magnitude

✉ Nicholas Emery-Xu
niemery@g.ucla.edu

✉ Richard Jordan
Richard_jordan@baylor.edu

Robert Trager
robtrager@gmail.com

[1] Department of Political Science, Baylor University, Waco, TX, USA

[2] University of California, Los Angeles, USA

[3] Oxford Martin AI Governance Initiative and International Governance Lead, Centre for the Governance of AI, Oxford, UK

Published online: 19 September 2024

more toxic than VX.[1] Most surprising of all, these researchers were not dedicated to uncovering novel toxins—they found them almost as an afterthought.

Consider this in the light of the ongoing revolution in large AI models like OpenAI's ChatGPT and DeepMind's AlphaFold. Future models built on these foundations will automate aspects of R&D processes across numerous military and civilian domains, rapidly uncovering wonders and new technologies of destruction. The 40,000 toxins of Collaborations Pharmaceuticals are the very tip of this iceberg, and no one knows where these developments will lead. What is clear is that technologies like this require governance—processes and policies—to ensure that AI systems are designed, developed, and used in a responsible and ethical manner. Political actors will also want to ensure that their interests are protected.

Policymakers have already begun to weigh the risks posed by transformative AI (TAI), or AI that will rapidly transform an existing socioeconomic domain. Antonio Guterres, Secretary-General of the United Nations, has created an AI Advisory Board to provide guiding principles on international AI governance. In testimony before the U.S. Senate, Sam Altman, CEO of OpenAI, called for tighter regulation of the industry; other industry leaders have called for a temporary moratorium on cutting-edge research. The UK and South Korea have convened Global Summits on AI in 2023 and 2024, respectively, with a future summit scheduled to take place in France in 2025. The G7 nations have created a "Hiroshima Process" to structure AI governance in the years ahead. Finally, a number of nations, including the US, UK, and Japan, have founded national AI Safety Institutes.

In this article, we analyze the broad scope of challenges and opportunities for governing transformational technologies such as advanced forms of AI. While AI has many meanings, we focus on machine learning. This AI technique uses (often large amounts of) computational power and data to train a model on prediction or classification tasks across a wide range of domains, ranging from predicting the structure of a protein molecule to identifying military targets. Modern AI systems are almost all based on the same set of machine learning algorithms and techniques, such that any sufficiently capable system is likely to have dual-use properties. In recent years, rapid progress in AI has led to predictions that it could irreversibly transform many existing socioeconomic and political domains.

Who will govern AI, and *when* will they intervene? We present a taxonomy of technology governance approaches along these two dimensions, who and when; this taxonomy clarifies what these different forms of governance can achieve and what historical precedents, if any, they imitate. We also introduce the concepts of the "supply" of and "demand" for governance at each stage of technology production processes; these concepts allow us to weigh how each governance approach can address classes of risks in different contexts. As appropriate, we discuss relevant international relations models and precedents for a variety of plausible structures.

Throughout, we will examine actors' incentives within a given structure, including incentives to free-ride, to cheat (e.g., violating thresholds on computing resources used to build an AI model), and to "cut corners" (e.g., reducing safeguards to accelerate development). As we discuss in the next section, the nature of transformative AI systems (TAI) presents different incentive structures from other recent transformative technologies like recombinant DNA, and any attempt to govern TAI must account for these differences.

AI governance has many goals, including stability, risk management, and the equitable distribution of the technology's fruits. All of these require controlling a technical process. They are also often linked: for instance, the nuclear non-proliferation regime both fosters global stability and redistributes the benefits of nuclear energy. The Hiroshima Process has, for the moment, privileged a "risk-based" approach to AI governance that seeks to make AI "trustworthy." How the international community will achieve such goals remains unclear.

We begin with subnational governance, which is prevalent and mitigates some risks. We argue, however, that subnational governance is insufficient when the individual rewards from taking risky actions outweigh normative sanctions. Similarly, nationally enforced standards are promising ways to govern AI *deployment*, but they are less viable in "race-to-the-bottom" environments surrounding AI development and proliferation. We contend, therefore, that some form of international governance will be needed.

When it comes to powerful technologies with military implications, a non-proliferation plus norms-of-use regime is the *only* multilateral option with a strong historical precedent. In plausible contexts, however, it is not effective on its own. Verification-based development and use restrictions, such as those advocated by the Secretary-General, would address a range of risks, but they face challenges in the case of advanced AI; moreover, there is no clear example of major powers restricting the development of a powerful military technology that does not have a substitute technology. International Monopoly, which might evolve naturally or be created, was the preferred solution of thinkers in the early nuclear era, and we dwell on how these midcentury proposals for nuclear weapons might be revived for TAI. A monopoly requires less-invasive monitoring, but at the possible cost of eroding national sovereignty. Overall, no single approach dominates the others.

Among international governance options for transformative technologies, we argue that only four are potentially

[1] Urbina et al. (2022).

viable when it comes to governing states' *military* sectors. Which of these four is best will depend on the technological and social contexts, and any eventual regime may include elements of several of these four.

I. A state-enforced non-proliferation regime with clear norms of use (NPT+).
II. An intrusive monitoring scheme extensively supervising AI development in all nations; this would take the form either of technology caps (if enforced by states) or International Verification (if enforced by an international body).
III. Deliberate internationalization of the technology such that a single actor or organization controls the cutting edge of AI technology (an International Monopoly).
IV. De-facto hegemony resulting from economies of scale, an ever-increasing technological lead, or locking in an (otherwise temporary) decisive strategic advantage.

Each of these is associated with a set of ethical concerns, and probably none more so than the last. Nevertheless, that option is perhaps not unlikely. The United States might have exercised such an option when it briefly monopolized nuclear weapons. We have no reason to believe TAI will be developed by an actor who would show the same restraint. Whatever rough magic is slouching to be born, it is uncertain whether their creators will abjure such potent art.

When nuclear weapons were invented, many believed they were a unique technological challenge. That view looks antiquated today. Technological progress appears to have sped up, and emerging technologies may demand governance solutions of similar scope to those proposed during the early nuclear era. That era broke upon an unprepared world. By contrast, we appear to have a window to consider the coming potential transformations, including how to shape technological and governance trajectories.

1 Risks and opportunities

Transformative technologies, like TAI, are technologies with the potential to produce a rapid, irreversible change in an existing socioeconomic domain.[2] Transformative technologies vary along two dimensions, depth and breadth.[3] Nuclear weapons were a deeply transformative but narrow technology, radically altering warfare but few other sectors. In its early years, information technology (IT) was a broad but shallow transformative technology, mildly increasing productivity growth across much of the economy and leading to the famous quip by economist Robert Solow: "You can see the computer age everywhere but in the productivity statistics."[4] In contrast, the rapidly advancing frontier AI appears likely to be transformative along both dimensions.

Scholars distinguish TAI from artificial "general intelligence" (AGI) and the often-nightmarish Skynets and cyborgs of popular imagination to consider how more incremental advances in AI might still upend existing institutions.[5] A RAND workshop suggests, for instance, that near-term AI capabilities could undermine nuclear stability, even between countries like the US and China.[6] Likewise, tomorrow's instruments of authoritarian repression are likely to include increasingly automated surveillance and social control technologies.[7] It is no longer controversial to suggest that technological developments of the coming years may lead to social upheavals and thus demand governance responses. Even the idea that AI pursuing goals misaligned with human values could lead to catastrophic risks for humanity has moved into the mainstream: The UK's National AI Strategy states that "the government takes the long-term risk of nonaligned AGI… seriously" and notes that such concerns are "by no means restricted to the fringes of the [computer science] field."[8] Indeed, AI *as it currently exists* may suffice to work an economic upheaval on the scale of the industrial revolution over the coming decades, transforming the global economic balance.[9]

Regulating such a still-emerging technology involves a timing problem known as the Collingridge Dilemma.[10] Regulate too early when risks are unclear, and one risks

[2] For another definition, see Dafoe (2018). There is necessarily some ambiguity in the definition of TAI. To make it more precise, some authors suggest a standard requiring TAI be "comparable" in its effects to the Industrial or Agricultural Revolutions, i.e., comparable to the invention of electricity or steam power. See Ross Gruetzemacher and Jess Whittlestone, "Defining and Unpacking Transformative AI" (unpublished manuscript 2019). We remain agnostic whether such standards ought to be adopted in scholarly or professional discourses.

[3] Gruetzemacher and Whittlestone (2022).

[4] Solow (1987).

[5] Scholars and industry professionals who focus on the risks posed by TAI are not necessarily downplaying experts' fears around AGI or other potential developments. Rather, they often stress how more limited technological developments might still have radical social consequences.

[6] Geist and Lohn (2018).

[7] Beraja et al. (2024).

[8] *National AI Strategy of the United Kingdom* (London: 2022). https://www.gov.uk/government/publications/national-ai-strategy/national-ai-strategy-html-version#our-ten-year-plan-to-make-britain-a-global-ai-superpower.

[9] Nichols Crafts argues that if AI raised the productivity of R&D in the economy, it would have similar effects to the First Industrial Revolution. Besiroglu et al. provide empirical evidence that current deep learning techniques could increase the productivity of R&D. Crafts (2021) and Besiroglu et al. (2024).

[10] Collingridge (1980).

a mismatch between regulations and the risks they were designed to mitigate. Regulate too late, and the technology is already in the hands of actors less willing and able to be regulated. The dilemma is especially pronounced for AI, where the technology and the uncertainty surrounding it are changing rapidly.

Progress in AI has moved more quickly than informed observers predicted even a short time ago.[11] It is possible such changes will threaten fundamental social parameters that produce a relatively safe and stable international society. In this section, we survey some of the most important risks and opportunities confronting AI governance.

When it comes to risk, Kissinger et al. (2021) get straight to the point:

> Throughout history, many technologies have been dual-use. Others have spread easily and widely, and some have had tremendous destructive potential. Until now, though, none has been all three: dual-use, easily spread, *and* potentially substantially destructive.[12]

Their characterization is not quite accurate—certain research agendas in, for instance, biotechnology would also seem to fit this description—but it is telling.[13] All of the qualities that make AI dangerous have been encountered before in other technologies, but rarely (perhaps never) all at the same time.

Internationally, states might attempt to prevent each other from developing technologies they consider harmful, but such regimes are likely to require intrusive monitoring of both small and great powers. What level of intrusive monitoring will states accept? The answer is unknown. The IAEA provides a model for monitoring dangerous technology, but one which applies largely to regional powers. Great powers have not submitted to similarly invasive processes. The problem is compounded when AI intersects cyber capabilities, because cyber weapons derive almost all their utility from their secrecy.[14] Counting does not diminish the number of explosives in a magazine, but an enumerated digital arsenal is a compromised arsenal. States have resisted such monitoring in the past on national security grounds, and they are likely to do so in the future.

Moreover, any regulation of advanced forms of AI will have to account for the billions of people who will interface with such a technology every day. Unlike nuclear technology, whose inputs are handled only by specialists, AI is in the hands of ordinary people, and it can be transported from place to place on ordinary devices. Like Chinese silkworms, AI is an extraordinary technology that lives in the most unextraordinary of places. To keep it out of reach of malicious actors or rival states, there is an argument for keeping advanced AI at arms' length from users, for example, so that users always interact through an Application Programming Interface or "API". This requires setting up regimes and norms *before* these technologies proliferate. Yet, it is unclear what kinds of denial and restriction strategies can succeed with such an everyday technology.

1.1 A typology of risks

When considering governance strategies, we can consider a typology of the risks they seek to combat. These risks can be classed as accident, misuse, and structural; and they might occur immediately or over time.[15] Accidental risks are those that arise as unintended consequences of technological development or deployment. Misuse risks are harms that arise from malevolent actors abusing a new technology. Finally, structural risks are those risks from a new technology that stem from changes in the surrounding environment or incentives.

Existing AI technologies already exhibit all three kinds of risk. AI algorithms, for example, have accidentally recreated deadly biological agents[16] and facilitated the dissemination of deepfakes and other misinformation.[17] AI labs, seeking to race ahead of their rivals in development, have reduced safety and ethical barriers to R&D, for instance by training models on improperly filtered datasets that result in outputs with racial or gender bias.[18]

In the longer term, AI's emergent features may drive additional grave risks. On one hand, diffusion of AI's economic benefits could be too slow: the data and computation intensity of current AI development could further concentrate market power as industries become more oligopolistic. On the other hand, diffusion of AI's risks could be too fast: if AI "supercharges" ordinary weapons by making them more destruction dominant, then any AI regime will need to prioritize non-proliferation.[19] Such weapons

[11] Zhang et al. (2022) and Grace et al. (2018)

[12] Kissinger et al. (2021)

[13] Stern (2003)

[14] For example. AI could be used to automate vulnerability discovery or phishing attacks. See Lohn and Jackson (2022).

[15] The distinction between structural risks from accident/misuse originates with Zwetsloot and Dafoe (2019).

[16] Urbina et al., "dual-use", 189-191.

[17] See Mirsky and Lee (2021) and Brundage et al. (2018).

[18] In response to the success of ChatGPT, Google has said it will "recalibrate" the level of risk, it is willing to accept in developing AI language models. Grant (2023). For an overview of some current risks of large language models, see Weidinger et al. (2021).

[19] The offense–defense balance is the ease with which an asset can be defended relative to the ease with which it can be taken. It has been extensively studied in the IR literature. Traditionally, scholars have argued that as the offense becomes easier, relative to the defense, a system will become less stable and more war-prone. See Jervis (1978a, b), Glaser and Kaufmann (1998) and Hopf (1991).

could cause massive destruction in the hands of small groups or even individual actors. They might also increase the risk of war between states.[20] In addition to these proliferation risks, AI poses significant dangers at the development stage, just like the Manhattan Project before it.[21] Development, if conducted in an unsafe way, or in an environment where dangerous inventions might spread beyond cordons, could jeopardize human security in significant ways, and, some believe, even on an existential scale.[22] This danger could take a variety of forms, from applications of AI to bioweapons development to the "misaligned power-seeking AI" of some concerned technologists.[23]

To date, attention among scholars and policymakers has largely focused on the top left of Fig. 1: on the risk of accidents (like racially-biased AI) or on the risk of misuse (like disinformation). This emphasis may lead to different approaches to AI governance than if analysts focused on the lower right. Zwetsloot and Dafoe write:

> This [emphasis], in turn, places the policy spotlight on measures that focus on this last causal step: for example, ethical guidelines for users and engineers, restrictions on obviously dangerous technology, and punishing culpable individuals to deter future misuse. Often, though, the relevant causal chain is much longer—and the opportunities for policy intervention much greater—than these perspectives suggest.[24]

As Zwetsloot and Dafoe recognize, including structural risk in the conversation requires treating AI governance as a fundamentally *political* question, rather than one of engineering or professional ethics. It also requires considering interventions at earlier stages of AI, i.e., at its proliferation or development stages rather than only its deployment. Fortunately, AI may have some qualities that *facilitate* governance rather than impede it.

1.2 Chokepoints and the Matthew principle

Unlike other transformative technologies, AI has two features that can be exploited to enable international governance: its supply chain has several chokepoints, and its production may tend toward market concentration.

AI can be thought of as having three principal inputs, known as the "AI triad": algorithms, data, and compute.[25] The first two are ephemeral. Software and data are relatively easy to copy, share, and steal. Software is also like speech and so difficult to regulate: its regulation could raise First Amendment concerns in the United States, and similar concerns elsewhere.[26] Setting aside such concerns, the capability to generate the algorithms required to produce cutting-edge AI is fairly widely dispersed, at least at the moment, making them hard to police.[27] And even were that capability not so dispersed, it is likely that AI algorithms could be stolen by determined actors. Thus, algorithms are difficult to regulate, and for similar reasons, data can be as well.

Happily, regulating TAI's physical infrastructure may be more feasible than for past transformative technologies like nuclear weapons.[28] The computing power required for many frontier applications of AI technologies is massive, currently on the order of USD 100 million. If some of the most significant risks are associated with the scale of computing power, as some analysts argue, there may be opportunities to limit the number of actors who can possess this capability and the ways they can use it.[29] There are relatively few providers in the world today. Further, the supply chains that produce the data center quality chips—chips that facilitate parallel processing with very high interconnect speeds—are narrow and exhibit high degrees of vertical integration, often with a single firm such as ASML producing a necessary component for the most advanced chips.

Thus, it makes sense to seek "chokepoints"—points where production processes require some controllable input—that can reduce a large portion of total risk. Since the training of AI systems today requires large amounts of computing power, one such chokepoint could be the monitoring of semiconductors (perhaps through chain-of-custody accounting from fabricators to data centers) and elements of the semiconductor supply chain. Other potential chokepoints could include the migration of top AI talent or preventing the sharing of state-of-the-art algorithms.

[20] Bas and Coe (2012) and Ben Garfinkel and Dafoe (2019).

[21] Ord (2020) and Stern (2002).

[22] Trager et al. (2022), Armstrong et al. (2016).

[23] Russell (2019).

[24] Zwetsloot and Dafoe, "Thinking about Risks from AI.".

[25] Compute refers to the computing power, usually graphical processing units (GPUs), used to train an AI model. Buchanan (2020).

[26] Speech and ideas are sometimes regulated, but such regulation is never simple. For instance, according to the "born secret" doctrine governing nuclear technology, the ideas associated with constructing nuclear weapons are classified even if they are discovered without the aid of classified sources. It is worth noting that this doctrine's constitutionality remains unsettled.

[27] This has occurred via the development of open-source alternatives and the leaking of source code. For an example of the former, HuggingFace's BLOOM emerged as a competitor to OpenAI's GPT-3. Scao et al. (2022).

[28] AI's compute infrastructure may be "governance-enabling." It seems likely that training AI will require increasingly specialized chips. Their highly specialized nature and supply chain make them a logical "choke point" for controlling compute, and thus TAI. Nonetheless, such control may presume that actors like China do not develop independent production capabilities. Khan and Mann (2020).

[29] Many properties of current language models can be predicted reliably from empirical scaling laws. Hoffman et al. (2022).

Fig. 1 A typology of risks

	Type of Risk		
	Accident	**Misuse**	**Structural**
Current	Racist AI, Self-driving car accidents	Electoral interference (e.g. via deepfakes or disinformation)	Rapid increase in energy use by big tech firms
Prospective	Misaligned, power-seeking AI	Killer robots, cyber worms proliferation, AI-enabled bioweapons	Increased risk of war; eroded MAD; labor displacement; monopolies and global inequality

Also, while AI models cost millions of dollars to train, they may require as little as a few cents per use to operate.[30] As a result, AI has significant economies of scale, and at least for now, its most transformative forms will likely require massive resources, resources only a large firm, and possibly only a large state, could muster.[31] In part, US and Chinese firms enjoy their lead over European and Southeast Asian rivals because of scale. Efforts to regulate domestic firms, or even to break them apart, could cripple these firms against their competitors abroad. Conversely, actors like the United States may derive significant political advantages when their firms gain oligopolistic power. Protecting citizens from predatory firms may come at the price of reducing international power.[32] On the other hand, to avoid this tradeoff, it may be possible to prevent an AI race by pooling resources in a single international body; this body might deter entrants not through coercion but rather through an unassailable technological lead.

Furthermore, once a firm develops a transformative technology, it may begin to pull *farther ahead* of its competition. Currently, the time to replicate for AI achievements is often measured in months, or even shorter. This may not hold in a world of TAI, where AI generates technological insights that compound on each other and can be used across many human domains. Indeed, if the best AI systems require training on large amounts of specialized data, the costs of acquiring such data could limit progress to a select few firms. Likewise, if firms only integrate their own AI systems into popular business or consumer software suites, such horizontal integration could limit TAI development to existing incumbents in the software industry. Here, if some firms or states begin to achieve a decisive lead, it may become ever more difficult for their competitors to close the gap: as consumers shift resources to the firm with the better product (a product which improves efficiency in many areas of life), the dominant firm will enjoy an increasingly unassailable position. If this happens, it might *enhance* the viability of some forms of international governance, because it can increase the "time to breakout" and deter entry even when monitoring and verification are less than ideal. AI may exhibit the classic economic tendency where "to whom much is given, more will be given."

These opportunities will be especially viable if states continue to lag behind firms in AI development. The potential Manhattan Projects of today are being executed primarily by non-state actors. Because firms are vulnerable to domestic and international law in a way states are not, widespread AI cooperation may be more feasible than for seemingly similar technologies, like nuclear weapons. States are notorious for breaking their treaties, but firms do not enjoy the same freedom. Even technological laggards can impose billion-dollar fines on major multinationals like Google.[33] Nonetheless, we note that the process of building an AI regime may unravel this opportunity: as states recognize the gap between themselves and firms, they may seek to close it.[34] It is not yet clear how stable firms' dominance of AI will be, once the technology becomes especially powerful, and states pay more attention.

2 Theorizing AI governance

We classify AI governance regimes along two dimensions, *when* and *who*. "When" sorts governance schemes by the stage at which they primarily try to limit AI: its development, proliferation, or deployment. The choice of "when" depends in part on risk assessment: those who see AI as an imminent or existential danger tend to advocate intervening at earlier stages of its development and production. Much like with nuclear weapons, those who most fear AI tend to favor restricting or even banning the means of its development,

[30] Altman (2022).

[31] Thompson et al. (2022).

[32] Kissinger et al., *The Age of AI*, 122.

[33] Chan (2022).

[34] This dynamic partly depends on the degree to which AI diffuses from states to firms. Horowitz (2018).

not just its use or proliferation. Currently, governance focuses mostly on how actors deploy technologies, for example protecting privacy or avoiding racism once a technology emerges. Yet, this focus on deployment may become untenable. As the AI firm DeepMind recognized when it released the protein folding prediction technology AlphaFold, "We must *proactively* evaluate the ethical implications of… research and its applications."[35]

Intervening at earlier stages also constructs a kind of defense-in-depth: the earlier governance can exert control, the more opportunities it has to head off a problem. All else equal, those who see AI as more dangerous will seek more lines of defense, and so will favor intervening as early as possible; by contrast, those who see AI as less potentially dangerous relative to its benefits will prefer to intervene only at the deployment stage, leaving its development and proliferation less inhibited.

The level of governance actors will accept also varies with their uncertainty over their relative position in technology races. This uncertainty affects the *when* of international governance because intervening at different points in production processes involves differently distributed costs. Rawls' veil of ignorance is a useful analogy here: when a person does not know whether she stands to gain or lose from a regime, she may prefer the regime that maximizes the welfare of the meanest citizen. Likewise, if a state or firm does not know whether it stands to gain or lose from an AI race, it will prefer a regime that limits such a race and the fallout to the losers. As this uncertainty decreases—for instance, as a firm becomes more confident it is likely to win an AI race—its preferred level of governance may decline, or it might prefer forms of governance that limit the actors who can join the technology club. This could mean limiting the availability of inputs into production processes.

Choosing when to intervene can often depend on technological necessity, especially upon opportunities for control over inputs. Because AI systems require vastly more computing power to train than to operate, the number of actors who could operate a system is much larger than the number with the resources to build it. Governance of a smaller number of actors is often simpler, and so targeting governance at earlier stages of the production process might be easier to enforce. These governance opportunities might be quite dramatic. For instance, extreme ultraviolet lithography machines are currently necessary to produce the most advanced chips, and these machines are currently produced by only one company, ASML; as a consequence, US export restrictions have targeted them (and other chokepoints like them) to limit advanced AI proliferation to China. Besides the obvious non-proliferation benefits, controlling inputs at these early stages also increases the scope of potential AI governance at later stages.

There are, thus, a variety of factors that influence the *supply* and *demand* for governance at different stages of the technology production process. An optimal governance regime requires finding an equilibrium between these forces. In AI governance, this could mean identifying a point where actors are willing to intervene based on their subjective assessments of the risks and rewards of the technology at its different stages, from basic science to conceptualization and development through potential proliferation and deployment, with an ability to intervene based on the potential to control inputs at that stage and other factors.

The *who* of AI governance depends upon the strategic environment: on the risks actors are incentivized to take, on the distribution of benefits from the technology, and on the off-the-path outcomes should governance fail.[36] We conceive of the *who* as the actors who primarily enforce the regime: substate, national, or international.

If risks are primarily driven by coordination failures, then governance should focus on coordinating actors' expectations around a self-sustaining Schelling Point. In such a world, norms would take on outsized importance as coordination devices, and substate actors would be ideally placed to govern some aspects of AI. Professional organizations could determine publication standards for dual-use technologies and auditing practices before deploying new AI systems. Industry professionals and professional organizations would be well-placed to develop and codify norms around publication, use, and best practices. International Soft Law regulation could allow scientists and officials to cooperate in international standard setting, as in international securities regulation.

Alternatively, should development proceed unevenly, first, movers might impose standards unilaterally. Sometimes, this applies to national actors. The United States and other nations influence global aviation safety standards, for example, by prohibiting flights into their countries by airlines that do not meet their standards. Downstream actors can then codify the resulting norms into national and international law. In such a coordination game, governance would seek to prevent parallel, fragmented normative environments; it would also seek to ensure that small states and technological laggards enjoyed adequate representation.

A Prisoner's Dilemma, by contrast, calls for different governance structures, and for different actors to take the lead in enforcing them. In a competitive environment, norms alone may not be sustainable, as firms and states deviate from cooperation to pursue the profit or power they expect TAI

[35] "How our principles helped define AlphaFold's release," September 14, 2022, https://www.deepmind.com/blog/how-our-principles-helped-define-alphafolds-release

[36] Jervis (1978a, b).

might bring. They might cut corners in safe development to outcompete rivals or even precipitate an arms race. In such a competitive environment, explicit rewards and sanctions may be necessary to deter actors from inefficient or unsafe competitive practices and to incentivize participation so that unregulated actors do not race ahead. Governance would require strong states and/or international bodies to enforce safe, beneficial development and to prevent any single actor from gaining a monopoly or hegemonic power.

In either world, a failure to govern would reduce global welfare. Miscoordination frictions resulting from AI developers complying with multiple sets of norms and standards would reduce economies of scale, as if two Internets operated in parallel. Racing dynamics could lead AI firms to launch unsafe or biased products to beat rivals to market. Unchecked competition and winner-take-all dynamics could exacerbate an unequal distribution of the benefits of AI, between firms who use their market power to stifle innovation and raise prices, or between states who reinforce or remake the balance of power. In competitive contexts like these, we argue that both national and international actors are likely to have important roles to play as enforcers of international governance regimes.

In this article, we survey tools, including National Standards and International Soft Law, for coordinating actors to maximize the benefits of AI. If coordination remains the primary obstacle to effective governance, then it is likely these tools will remain vital to ensure the economic benefits of AI are maximized and equitably distributed. On the other hand, we argue it is more likely states will compete to harness AI to maximize military and economic power. First, as current AI is nearly exclusively built upon deep learning, economic innovations are likely to rapidly diffuse into advances in military technology. Second, should TAI rapidly automate large segments of the economy, states may seek to exploit the resulting increased economic growth to enhance their own power. Thus, any discussion of international AI governance must account for its security implications as well as its economic ones.[37]

Figure 2 summarizes the governance options available based on these two dimensions. For simplicity, we break the when of governance into three stages: development, proliferation, and deployment. We now turn to analyzing the benefits and drawbacks of these governance options. We consider each of the three sets of enforcing actors in turn.

3 Subnational governance

Who will govern AI? We first examine subnational groups. They, not states, are the most obvious candidates. Indeed, AI is rapidly leaving states behind. Its complexity, and the speed at which it is advancing, make the trundling terrapin of modern bureaucracy ill-suited to keep pace with it. AI's complexity suggests the need for expert governance, governance that values competence above all else. Its speed of development suggests the need for guardrails rather than supervision, for rules that set boundaries and otherwise get out of the way. Firms, nonprofits, and professional organizations are, therefore, well-placed to take the lead—as indeed they already have. To date, much AI governance has come from subnational actors; this fact demands we take them seriously as *the* potential anchor for any regime.

Forms of subnational governance, such as publication norms, can have dramatic effects. In 1940, for instance, Leó Szilard convinced Louis Turner not to publish the idea for a plutonium bomb—which might have radically altered world history.[38] More recently, following years of frustration with the slow speed of government regulation, private groups are engaging in norm development and service provision usually associated with states. A particularly striking example is the International Biosecurity and Biosafety Initiative for Science (IBBIS), which proposes to screen DNA synthesis orders to prevent misuse of diffusing biosynthesis technologies. As the organization points out, "94% of countries have no national-level oversight measures for dual-use research, no agency responsible for such oversight, and no evidence of national assessment of dual-use research."[39]

These strengths of subnational governance are visible in the ways professional organizations shepherd AI proliferation. Researchers are often best-positioned to understand the risks a technology may pose. Even though the boundaries of acceptable research are relatively undefined, many AI professionals are engaged in a continual dialogue about recent achievements. In its early days, AI enjoyed a startling openness with its discoveries, with most working papers and replication code publicly available. Over the past few years, rising concerns about the potential misuse of its inventions have diminished this transparency. Norms are evolving daily to govern just how accessible researchers should make their creations. When a new discovery becomes problematic, these social networks rapidly raise awareness about the ways malicious users can (and will) abuse it. For instance, facial recognition technology has become a tool of

[37] US semiconductor export controls on China, for example, show that nations are keenly aware of the security implications of AI technologies.

[38] Ord, *Precipice*, 32. Building a plutonium bomb could have been attractive to a range of powers, including Germany, since it did not require isotope separation.

[39] Hamburg et al. (2022).

political repression,[40] and natural language programs have spread bigotry and misinformation.[41] As a result, firms and researchers have taken steps to prevent these technologies from falling into malign hands.[42] Once a discovery is made, the wider profession has shown a marked ability to shape and restrict its dissemination.[43] Similarly, thousands of Google employees protested when the company participated in a Pentagon program that could be used to improve the targeting of drone strikes.[44] Seemingly in response, the company promised that it would not develop artificial intelligence for weapons technologies, and it let the project lapse.

3.1 The Asilomar model

The celebrated Asilomar Conference on Recombinant DNA is often held up as a model for subnational AI governance. This example from another realm of dangerous, complex, and rapid innovation offers much to recommend and much to approach with caution. In 1975, a carefully chosen group of scientists met in Asilomar, California to guide future research around recombinant DNA. They sought to distill broad disagreements about appropriate norms in their field into tangible consensus policies. Without a united front, as the Nobelist David Baltimore noted at the time, “we will have failed in the mission before us.”[45] In advance of the conference, a strict moratorium was imposed on further research.

To a large extent, the Conference participants did agree to a set of measures to contain risks. This consensus then shaped subsequent field standards.[46] Far from stopping research in recombinant DNA, however, the Asilomar Conference facilitated it, albeit within certain guidelines and alongside new safety measures. Indeed, scholars in the field would not have accepted greater restrictions, and some resented those that were put in place.[47] And the moratorium Asilomar imposed was necessarily and strictly temporary: from the beginning, the goal was to put guardrails around future research, *not* ban it.[48] We should probably expect that other regimes in the Asilomar mold would behave likewise.

Part of Asilomar’s enduring appeal is the way it privileged expertise. Asilomar strove, as much as possible, to keep governance “in-house.” It empowered experts to restrain each other from potentially dangerous research agendas, and it did so without compromising the scientific process. Indeed, it offers a compelling playbook to insulate research from know-nothing politics. Because its architects felt that public opinion should inform, not dictate, its standards, Asilomar forewent legal enforcement unless absolutely necessary. It, therefore, tried to limit governance as much as possible to technical issues, since doing so would enhance professionals’ legitimacy when enacting and enforcing a largely nondemocratic regime that excluded many potential decision-makers.

3.2 Subnational governance is insufficient to address TAI

The Asilomar conference shows how subnational governance might avert catastrophe. Its example seems especially valuable “when risks are unproved and uncertain:”[49] it does not foreclose potentially valuable discoveries, but instead offers a way to progress while avoiding the most dangerous mistakes. Yet, Asilomar also demonstrates the weaknesses of a subnational regime. It failed to prevent many instances of misuse, including gene-edited babies.[50] More generally, Asilomar-style governance requires a consensus on standards, since social sanctions will fail without broad buy-in across a given field. As well, while justly celebrated, the original Asilomar conference and its aftermath are also much maligned, precisely because the broader public had comparatively little sway over the final regime. Insulation from political accountability can diminish the responsibility experts feel to their fellow citizens and even deafen them to their just concerns. Finally, Asilomar must rely on professional rewards and punishments to control researchers’ behavior; it offers no way to restrain experts whose rewards are not professional or who operate outside its social circles.

Despite these limitations, subnational governance of transformative technologies may enjoy future successes where

[40] Beraja et al. (2023).

[41] Weidinger et al. (2021).

[42] A notable move in this direction was IBM’s refusal to pursue facial recognition technology.

[43] Note: while publication norms enjoin experts to consider the “downstream consequences” of their research, they primarily aim to limit *proliferation*, not the research itself.

[44] Shane and Wakabayashi (2018).

[45] Rodgers (1977).

[46] Fredrickson (1991).

[47] Ephraim Anderson, for example, a British bacteriologist who had attended the conference and previously taken issue with the “pompous” nature of the earlier moratorium, reflected on the event critically: The meeting had reminded him of “Bernard Shaw's definition of the English gentleman hunting the fox: the unspeakable in pursuit of the uneatable…here was a bunch of people, with no experience in the handling of pathogens, virtually, with the sole exception of a mere handful, considering hazards that were not even known to exist.

Footnote 47 (continued)

There's a certain comic atmosphere about it.” Frederickson, “Asilomar.”.

[48] Krimsky (1982). It is worth noting that the scientists at the Asilomar conference, and in the profession at large, were the ones most determined to end the moratorium.

[49] Krimsky, *Genetic Alchemy*, 64.

[50] Cohen (2018).

When do they govern AI? \\ Who enforces the regime?	Subnational	National	Supernational
Development	**Asilomar** Professional organizations establish clear guardrails on R&D, socialize young professionals into these norms, and ostracize those who violate them. Historical Precedents Asilomar II (Recombinant DNA) ICANN AI Proposals Campaign to Stop Killer Robots	**National Verification** States impose technology caps; they restrict races within verifiable guardrails, enforce these standards via domestic law and/or export controls on inputs, and retaliate against/refuse to cooperate with states who do not enforce similar standards. In extreme cases, these guardrails could ban certain technologies. Historical Precedents Biological Weapons Convention	**International Verification** An international agency monitors development. It has the capacity and authority to ensure compliance. Historical Precedents: IAEA, Chemical Weapons Convention AI Proposals An international AI agency similar to the IAEA; this IAIA might monitor global chip supply chains and/or all AI training runs above a certain compute threshold
Proliferation	**Publication Norms** Professional/industry organizations monitor advances and develop norms governing technology and knowledge spread Historical Precedents Consortium for Monitoring, Technology, and Verification	**Non-Proliferation Regime** States agree not to spread a technology, and they employ carrots/sticks (such as export controls) to prevent its spread to those outside their "club." Historical Precedents NPT, JCPOA, Australia Group AI Proposals Export controls	**International Monopoly** A single actor maintains exclusive control of a dangerous technology and deters entry by other actors; such a monopoly could arise accidentally or be designed prospectively. Historical Precedents Acheson-Lilienthal Proposal (failed) AI Proposals CERN for AI
Deployment	**Soft Law** Experts propagate values and best-practices for new technology. Historical Precedents Professional association standards AI Proposals Standards proposed through the Partnership on AI	**National Standards** States enact standards (incl. licensing/cert requirements) governing technology use. Ideally, these standards would be harmonized across like minded states. Historical Precedents OECD Privacy Principles, commitments around WMD first use, NIST, some arms control AI Proposals EU AI Act, NIST Standards	**International Regulation** An international organization regulates the deployment of technologies. These regulations might be incorporated automatically into domestic law. Historical Precedents WTO

Fig. 2 A taxonomy of AI governance

it has, in the past, come up short.[51] Professions and academic disciplines are increasingly dense, in part thanks to social networks and AI. Widespread coordination and normative entrepreneurship are becoming easier and easier. It seems reasonable to expect that social sanctions will likewise be easier to enforce around the development of AI than they were around previous technologies. For instance, publication norms and norms against cooperating with defense establishments may become stronger. Alexander Grothendieck, a Fields Medalist and one of the most consequential mathematicians of the twentieth century, urged his colleagues to impose more severe social and professional costs against colleagues who cooperated on weapons production; ultimately, he seems to have left the field in part because of concerns about his work's applications to defense. (Later, in a Paris lecture, he railed against the way his discipline only "softly" distanced itself from those who cooperated with defense establishments.)[52] Such social sanctions will likely be increasingly viable in the years ahead. If so, then we should expect that governance

[51] It is worth noting that nonstate actors play an essential role in climate governance as part of a "hybrid multilateralism;" see Mayer and Zahar (2021), Abbott (2012) and Partzsch (2020). Some authors even see them as the primary agenda-setters, though this view is controversial. Whatever the case, the significant role of nonstate actors in ongoing debates about climate governance suggests that they could enjoy a similarly prominent role in debates about AI governance.

[52] "this minority [of scientists who cooperate with defense establishments] is more or less disavowed… rather softly…. Far from banishing them from the scientific community, the fact that a scientist collaborates actively with the military does not in any way prevent him from fulfilling important functions in any scientific society, nor from having cordial relations, even friendly with most of those of these colleagues who, on their own account, have objections to active collaboration with the army." Alexander Grothendieck, "Responsabilité du Savant dans le Monde d'Aujourd'hui. Le Savant et l'Appareil Militaire.".

regimes relying on social enforcement mechanisms will be more successful and more numerous than in the past—assuming they can reach the necessary consensus.[53]

Still, we should not expect too much from social pressures, even in the age of social media. The "move fast and break things" ethos of Silicon Valley often causes ethics to fall by the wayside.[54] Moreover, nationalism tends to trump other forms of idealism, and we should probably expect that, in the shadow of international tensions, professionals will close ranks with their neighbors, whatever the social norms in their fields. In the years before WWI, it was hoped (or feared) that the "workers of the world" would truly unite and refuse to fight for their countries. In the event, nationalism overwhelmingly trumped class solidarity, and the same people who marched for international solidarity on May Day marched for the Emperor just 3 months later.[55]

In short, subnational governance is well-suited to influencing the rate of proliferation and the direction of AI development at the margin. It cannot, however, withstand strong structural pressures, if for instance the first state to develop a technology gains a significant political advantage, or if espionage is an especially effective tool to close technological gaps. If a technology would significantly shift the balance of power between states, or if it would fabulously enrich the firm that first invents it, then a subnational regime's enforcement mechanisms will prove too weak to govern AI.[56] It may, therefore, be best-suited for governing modest advances in industry (like more accurate facial recognition software) rather than the development/proliferation of transformative technologies.

For these reasons, it should perhaps be unsurprising that the Asilomar Conference on Beneficial AI, which tried to replicate the success of its biological predecessor as literally as possible—perhaps a touch too literally—fell short of its predecessor. It articulated a set of principles for AI development, but it failed to adopt concrete measures to contain risk. Subnational actors are most effective at governing technologies when i) there is consensus around the appropriate ends of governance and ii) the threat of social sanction is not overshadowed by other imperatives, such as actors' drive for security or profit. In other words, subnational governance is most effective when supervising technologies with modest security implications or when political actors are paying little attention. Today, states are paying ever more attention to the development of AI, and the technology is likely to have large impacts on corporate profits and on national security. Therefore, while it is difficult to know all the areas where subnational governance is likely to fail, it seems highly likely that it will fail in some aspects of governing transformative AI.

4 National standards

Fundamentally, if AI threatens to reshape the international economy or the balance of international power, then the *who* of AI governance will be a state or an international body, at least in significant part. Already, the European Union seems to have concluded that states and superstates should have the whip hand in AI governance. The United States is, to date, more ambivalent, but statesmen like Henry Kissinger seem to take eventual "nationalization" of the AI problem for granted.[57] Major powers will need to limit their strategic dependence on adversaries, counterbalance against changes in international power, and protect their citizens from exploitation by foreign actors. States also seem to be realizing that no other actors are or will be in a position to negotiate around such a widespread, politically and economically transformative technology.

4.1 National standards are promising ways to govern AI deployment

Moving from substate actors to state-based ones risks compromising the role of experts in any AI regime. One of Asilomar's virtues was its insulation of science from political pressures. A National Standards regime tries

[53] Already, top AI conferences such as NeurIPS have introduced required impact statements for paper submissions, Prunkl et al. (2021).

[54] Both Microsoft and Google have fired members of their AI ethics teams in recent years as competition over the technology has intensified. Schiffer (2021) and Schiffer and Newton (2021).

[55] One historian describes the before-and-after in Vienna: "On May Day 1914, workers had marched on the Ringstrasse with the chant '*Frieden, Brot, und Freiheit!*' ('Peace, Bread, and Freedom!'). On August 1 [after the assassination of Franz Ferdinand], many of the same crowd marched again with 'Alle Serben mussen sterben!' ('All Serbs must die!')." Morton (1989). More broadly, the distinguished military historian Michael Howard calls it "the most remarkable phenomenon of 1914–the excited crowds filling boulevards of every major European city…the masses of men required by military professionals came forward with superabundant goodwill." Howard firmly rejects the idea that this enthusiasm was the result of elite manipulation, indoctrination, or propaganda. Rather, nationalism was simply more appealing than alternative ideals. Howard (2009). Finally, it may also be worth noting that Kenneth Waltz' *Man, the State, and War*, which some would name the foundational text of modern IR scholarship, devotes an entire chapter to the failure of the Second International and the triumph of national interest over international solidarity among the proletariat in 1914.

[56] For surveys of the effects of military innovation on the balance of power, see Grissom (2006) and Horowitz (2020). See, in particular, Biddle (2004), Horowitz (2010), Gilli and Gilli (2018/19).

[57] Kissinger et al., *The Age of AI*, 119, 128, 172, 216, 224.

to harness the coercive tools of modern states while still leaving rulemaking itself in more-or-less professional hands. It is essentially a *transgovernmental* approach.

In a transgovernmental regime, standards are enforced by national governments, but these standards are primarily articulated by mid-level officials without the direct involvement of national political figures. These standards are often then harmonized across like-minded states, for example central bankers from Europe and America harmonizing capital requirements.[58] Transgovernmentalism has become a common way to embrace expert governance while maintaining national sovereignty, two goals that are especially valuable when an issue is both highly technical and politically significant, like banking and TAI. It is also an integral part of any "defense-in-depth" against a complex risk environment, especially when risks may require multiple kinds of responses; in this environment, National Standards can bridge governance efforts at the subnational, national, and supranational levels.

Where subnational models envision a primarily self-enforcing set of standards among professionals, the National Standards model insists that states make these standards legally enforceable. Their articulation as, for example, NIH guidelines is insufficient. Certainly, standards that are not legally binding can be efficacious in some situations, but they then generally fall under subnational governance in the form of "soft law" because they are enforced through peers' social sanctions rather than by the state. Naturally, it follows that binding National Standards will empower different experts than an Asilomar-style regime, viz. experts within regulatory bodies rather than experts in industry and the academy.

The National Standards Model revolves around propagating good ideas, improving legal instruments, and harmonizing standards across like-minded states. The model, therefore, recommends regular contact between similarly situated regulators and academics as the best means to govern a transformative technology. Already, some nascent AI governance follows this mold. Organizations like the National Institute of Standards and Technology (NIST) promote transnational standards and work with governments (including the US, via the National Artificial Intelligence Advisory Committee) to implement them. More aggressively, the proposed EU AI Act would severely restrict some forms of AI and closely regulate others.[59] In these efforts and others, National Standards closely track governing patterns that have emerged in the "new interdependence" of twenty-first century globalized spaces (Farrell and Newman 2016). These governance patterns enable close cooperation between international, state, and nonstate actors, and they often significantly empower nonstate actors to shape final rules and their implementation (Mitchell 2022). Such relatively low-cost approaches to international governance also seem to be moving the world away from formal treaties between states and away from the creation of new, large international institutions (Abbot and Faude 2021). If anything, this kind of governance tends toward "minilateralism," an approach now advocated by many climate activists.[60]

Like CO_2 emissions, AI research is concentrated in relatively few hands. National Standards seek to capitalize on this concentration. Rather than pursue a global solution from the beginning, National Standards can begin with a limited number of players. Imposing standards within these nations, and then gradually harmonizing them between these nations, might thus create "stepping stones" to a broader regime.[61] And just as with CO_2, the smaller steps of this more-targeted approach might, its advocates hope, be a more reliable way to eventually achieve a fully international solution.[62] On the downside, this approach is exclusionary, and the AI have-nots might find themselves left out of crucial decisions while such a regime takes shape.

4.2 National standards are less viable in "race-to-the-bottom" environments

Harmonizing standards across nations is often difficult to achieve. The National Standards model must, therefore, be prepared to accept that not all actors will play by the same rules. When it comes to deployment-stage challenges, this seems a small cost to pay: if France protects its citizens' privacy, and America does not, well, so much the worse for America. But this lackadaisical attitude becomes dangerous when confronting development and proliferation challenges, especially ones relevant to military technology.

In this vein, it is noteworthy that the National Standards model is pioneered by an AI laggard: with a few notable exceptions, European AI trails a distant third behind American and Chinese, and Europe shows few signs of closing the

[58] Slaughter (1997).

[59] It is worth noting that the act, which may be the boldest so far of its kind, mainly regulates the *deployment* of certain kinds of AI.

[60] Eckersley (2012)

[61] For a discussion of norm diffusion in the liberal school of international relations, see Ikenberry and Kupchan (1990).

[62] On the ways extended multilateral governance has become hidebound and ineffective, see Hale et al. (2013).

gap.[63] While it would be incorrect to attribute this gap to the European governance model, National Standards could hinder European firms' attempts to catch up.[64] And that speaks to a larger problem. If nations can gain by cutting corners in an international AI race, they may do so.[65] Restricting one's own country, without integrating those restrictions within a worldwide framework, is not unlike hoping unilateral disarmament will lead to world peace.

Nevertheless, like subnational governance, National Standards have an important role to play alongside other governance strategies. They mitigate risks from firms, though they are hard to enforce against secretive defense establishments. Indeed, sometimes National Standards are written with national security exemptions—for instance, prohibiting biological and chemical weapons research unless it is carried out in the service of the national government for purposes of developing counter measures. National Standards are more effective when countries that are developing the technology are aligned and, thus, less prone to violate effective standards in the name of national security. For example, while Chinese government bodies and think tanks have published white papers outlining governing principles for ethical and transparent AI algorithms, in practice the government has used AI to further domestic surveillance.[66] This implies that the effectiveness of National Standards approaches may be contingent upon non-proliferation regimes.

5 The NPT+ option

If we assume that AI could dramatically shift the international balance of power, then National Standards or Asilomar-type approaches, while helpful at the margin for mitigating risks from firms, may not effectively address risks that derive from intergovernmental competition. In this environment, unless a single actor can prevent progress elsewhere, an agreement among states—formal or informal—may be the minimally necessary foundation for an AI regime.

5.1 Non-proliferation has strong historical precedents

In this section, we consider what we call the "NPT + Model" for governing a transformative technology. NPT + takes the Nuclear Non-Proliferation Treaty as its touchstone. It envisions a limited competition among major powers, a "club," with access to a transformative technology strictly limited beyond their number. In the case of TAI, it centers AI governance at the non-proliferation stage. Doing so reduces the need for invasive verification schemes (see below). Likewise, because it leaves wide latitude for a few great powers, NPT + sidesteps any need to aggressively limit the physical AI infrastructure nations like China or the US might seek to create.

NPT + generally has three components: a non-proliferation regime to limit development to a few actors; a set of agreements or norms to govern the technology among the few who have it; and an economic development component to compensate actors who are denied the underlying technology. For AI, an NPT + regime might also facilitate broad access to the fruits of the technology through APIs and other forms of structured access.[67]

Non-proliferation regimes can be unilateral, multilateral but informal, or treaty based.[68] The recent Chips Act is a largely unilateral form of non-proliferation. It relies on export controls backed by US government carrots and sticks. The Americans are able to exercise this kind of unilateral governance by exploiting the dependence of companies in the semiconductor supply chain on chip design software from the United States.[69] Companies that violate the Act's restrictions would not be able to employ the design software. It is unclear how effective the Act will be over the medium-term as other actors develop alternative software.

The Wassenaar Arrangement is an example of a multilateral non-proliferation regime that is not codified by treaty. This group of 42 states exchanges information and polices the transfer to non-participating states of a wide range of weapons and weapons precursors. Despite being informal, Wassenaar is generally regarded as very successful. Other examples include the Australia Group, the Nuclear Suppliers Group, and the Missile Technology Control Regime (MTCR).

The Treaty on the Non-Proliferation of Nuclear Weapons (NPT) is the primary example of the treaty-based regime. Formal treaties have a variety of benefits, including political and reputational costs to states for violating them. They can also help states provide credible assurances, which are just

[63] Authors differ in how they identify the key firms in AI. Amy Webb's "big nine" are probably the most common (Google, Amazon, Apple, IBM, Microsoft, Facebook, Baidu, Alibaba, and Tencent). However, they are tallied, virtually all of the major firms are American, Chinese, or British. Of Webb's nine, all are American or Chinese. Webb (2019).

[64] Indeed, French AI firm Mistral used such an argument in an attempt to lobby European regulators to relax regulations on general-purpose AI systems. Bertuzzi (2023).

[65] Emery-Xu et al., Op. cit., 2022; Stafford et al., Op. cit., 2022. Such risks range from advancing performance of AI systems over their safety to the more traditional inefficient arms building up resulting from a security dilemma (Fearon 2011).

[66] For views on Chinese AI regulation, see "How will China's Generative AI Regulations Shape the Future? A DigiChina Forum," (2023).

[67] Shevlane (2022).

[68] For the classic view on arms control, see Schelling and Halperin (1961) and Bull (1973).

[69] Allen (2022).

as significant as credible threats but far less appreciated.[70] Their strictures, though, are not confined to the club they create. Treaties can impose obligations on actors excluded from the club: under the NPT, for instance, states that do not possess nuclear weapons technology commit not to develop it and submit to invasive inspections. In return, signatories without nuclear weapons receive assistance with their peaceful nuclear energy programs. Formal treaties are difficult to achieve, however, particularly in recent decades among the major powers. The Conference on Disarmament and its predecessor bodies, for instance, achieved major pieces of arms control legislation every few years for decades; it has achieved none for more than a quarter-century (Fig. 3).[71]

All non-proliferation regimes face three obstacles. First, actors may seek to erode them over time. Because the technologies they restrict are dual-use, those excluded from the club will pay significant economic costs. These actors, if not appropriately compensated, will protest. But it is not just those outside the club who may seek to erode the regime. Powerful states and their firms who would like to sell the technology to excluded actors will also seek to weaken the regime. This pressure will intensify as the threat environment, from which the regime derives its *raison d'etre*, diminishes. Thus, the strict Coordinating Committee for Multilateral Export Controls (CoCom) was succeeded by the weaker Wassenaar Arrangement following the Cold War's end. Paradoxically, non-proliferation may require *more* maintenance as the risk of catastrophe shrinks.

Second, powerful states may shirk their responsibilities in upholding the regime. Major powers have an incentive to provide security technologies to friendly states and to look the other way when those states violate the regime.[72] Since the regime relies on these powerful states for its enforcement, their willingness to flout its rules can jeopardize the entire structure.[73] Relatedly, major powers have incentives to stop providing other actors with positive and negative inducements to comply with the regime. Arms control agreements are frequently violated because actors do not invest enough in upholding them. The NPT regime functioned because it was a pillar of major power foreign policy—indeed, leading scholars have argued that the United States prioritized non-proliferation on the same level as European reconstruction and the Cold War balance of power.[74] These powers provided security guarantees to states who complied, and they threatened states who did not. During the latter half of the Cold War, they also divided the world into spheres of influence, which reduced conflict between those with and without nuclear technology. As this structure deteriorates, states like Ukraine may press to acquire such technologies themselves.

The third and final obstacle to non-proliferation is the challenge of monitoring.[75] The severity of this challenge is determined by the current state of technology, and this can be expected to change over time.[76] It is also a function of states' perceptions of how technological inputs translate into state power.[77] The proliferation of nuclear technologies and their inputs is probably easier to monitor than the proliferation of AI technologies, with the possible but important exception of the chips themselves (see above).

5.2 Non-proliferation has lower verification requirements and reduces risk in many scenarios

In spite of these challenges, non-proliferation regimes tend to be much easier to institute than other forms of multilateral governance. Often, they need not involve the consent of some affected actors, as export restriction regimes like the MTCR demonstrate. Sometimes, they do not need the full consent of *any* other actors, as in the case of the Chips Act, which the US unilaterally imposed. Non-proliferation regimes can also attract the support of powerful states more easily than other multilateral governance, because they reinforce rather than compromise these states' privileged technological positions.[78] By contrast, powerful states tend to fear regimes that limit their power—such as Technology Caps and International Monopolies—in case their rivals find ways to cheat or capture these institutions. Finally, in a non-proliferation regime, there is also less incentive

[70] Thomas Schelling originated the idea of a credible assurance. For recent work, see Cebul et al. (2021).

[71] Many recent arms control agreements, such as the Treaty on the Prohibition of Nuclear Weapons, the Anti-Personnel Mine Ban Convention, and the Convention on Cluster Munitions, have not included major powers.

[72] On the other hand, powerful states might be incentivized to seek arms control with another state to drive a wedge between it and a rival power. Crawford and Vu (2021).

[73] For example, the two superpowers colluded to uphold the nuclear NPT during the Cold War, but that may not be as viable in a multiploar world. Coe and Vaynman (2015).

[74] Gavin (2015).

[75] Occasionally, states such as South Africa have been able to develop nuclear weapons entirely undetected. Narang (2016). More generally, studies have found that the inability to perfectly monitor arming is a primary driver of arms races. Meirowitz and Sartori (2008).

[76] It may also depend on changes in the international system, as states without AI-Power security guarantees or facing strong domestic pressure could have high incentives to acquire powerful AI. Sagan (1994) and Solingen (2009).

[77] Logan (2022)

[78] Matthew Kroenig finds that Great Powers are more likely to limit proliferation, even among allies. Kroenig (2014).

to cheat because doing so has a smaller impact on the balance of power. The tenth state to acquire nuclear weapons changes the balance of power far less than the first. Major powers prefer international agreements that, if suddenly disrupted against them, will not leave them at a decisive disadvantage.[79]

Besides being easier to institute, non-proliferation is often *easier to monitor*, as well. It tends to require less-invasive methods than other forms of multilateral governance. Tracking the movement of goods between states is less invasive than tracking what is done within them. For advanced AI, this will likely be true in the case of hardware, and may even hold for software. It also makes it easier to build trust among the major powers: violations are easier to detect and attribute given the limited number of suspects, and norms of use are easier to create and sustain for the same reason. Indeed, non-proliferation might require only modest disclosures among the major powers, like states' military doctrines,[80] which unlike the capabilities themselves might potentially be revealed and scrutinized without compromising the underlying arsenals. And of course, when there are few competitors, information on each others' activities is also easier to obtain through espionage.

For these reasons, non-proliferation should be seen as a complement or *alternative* to a verification-based regime. It is built to circumvent the need for invasive verification. Nevertheless, monitoring remains a stumbling block even for well-designed non-proliferation regimes. Technologies and institutions that facilitate it will, therefore, make non-proliferation more feasible. For instance, a chip accounting regime,[81] or placing GPS trackers on chips, might greatly facilitate a non-proliferation regime that sought to limit concentrations of computing power outside the major powers.

While easier than other forms of multilateral governance, non-proliferation is still not easy in any absolute sense. When it comes to powerful, useful technologies, success will require non-proliferation to be a pillar of major powers' foreign policies. It will not succeed if they do not prioritize it. In this, it may be inferior to Asilomar-type regimes in which nonstate actors can take the lead.

In addition, NPT + could easily bifurcate international society into "haves" and "have-nots." It would entrench the divide between great and small powers. Whether by design or by accident, its namesake has clearly created a nuclear "club" whose members play by different rules than the rest of the world. This downside could be mitigated by guaranteeing the "have-nots" access to the economic benefits of TAI, in the same way that IAEA extends the benefits of nuclear technology to states without nuclear weapons.[82] Nonetheless, if transformative AI is entrusted to only a subset of nations, as it must be under NPT +, there is probably no way to avoid a stratified international system.[83][84]

5.3 Non-proliferation is most effective when risks plateau

The desirability of non-proliferation for transformative AI may boil down to three questions, none of which is yet answerable. First, can the benefits of the technology be enjoyed without transferring either the technology itself or those of its inputs associated with major risks? If not, countries and their citizens will be loath to tolerate restrictions, which would entail severe economic costs. For instance, if the same chips that are the lifeblood of national economies are sufficient for advanced AI, an NPT + regime restricting access to chips will be difficult or impossible to construct.[85]

Second, will AI facilitate destruction-dominant technologies, ones that can cause large-scale destruction against which there is no defense? If it does, it is probably imperative to keep those technologies in fewer hands, especially if small groups or even individual people could deploy them. NPT + would then be a necessary part of any AI governance regime. Non-proliferation regimes that reduce access without fully restricting it can also do more harm than good. For instance, a state with lower computing capacity may be *more* likely to take risks in the face

[79] On the other hand, if states have few defensive substitutes, they might still seek to acquire defense-biased AI technologies. Narang, "Strategies.".

[80] Kissinger et al., *The Age of AI*, 173–174 make a similar suggestion.

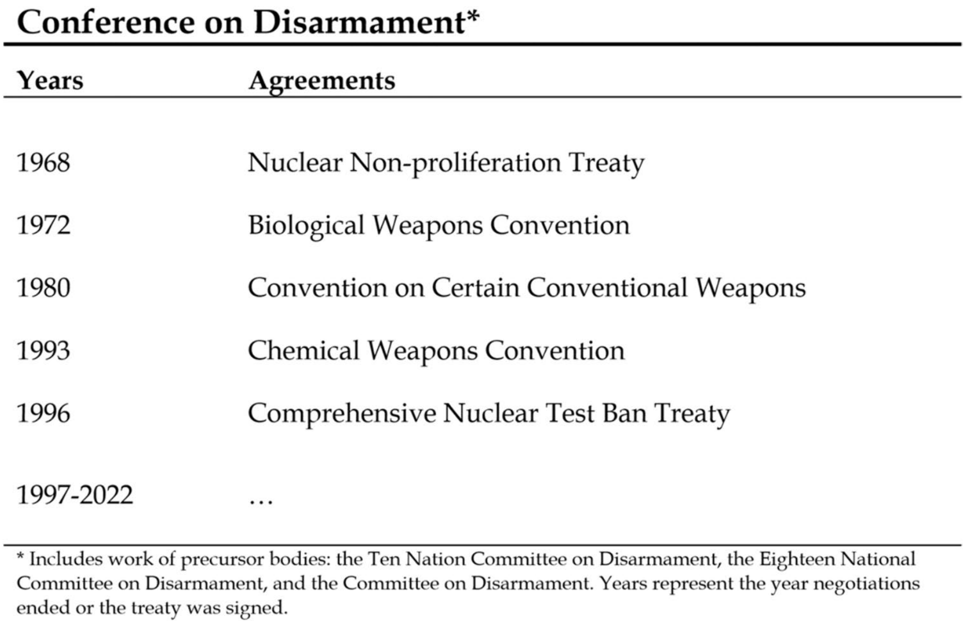
[81] A chip accounting regime would keep a "dynamic ledger" of all chips, or some other indispensable physical component, used in certain kinds of AI training runs. Given the chokepoints throughout the supply chain for advanced semiconductors, some analysts believe that such a ledger might be surprisingly feasible. Google has already patented a means of verifying unique chips.

[82] Indeed, formal guarantees to redistribute benefits can induce some less-powerful states to drop out of races for transformative AI, lowering the potential for a race to the bottom. Stafford and Trager (2022).

[83] It should be noted that AI's economies of scale have already left small nations at the mercy of larger ones and research universities at the mercy of large firms, and that trend will likely continue, whether or not international law creates a distinction. Ahmed and Wahed (2020).

[84] In such a world, security guarantees may also be required along with economic transfers to reduce states' fears about relative losses. Snidal (1991).

[85] For example, if advanced AI systems can be run on relatively few specialized chips such as GPUs or TPUs, then states may be able to substitute them for less effective CPUs, which, however, are far more important economically.



Conference on Disarmament*

Years	Agreements
1968	Nuclear Non-proliferation Treaty
1972	Biological Weapons Convention
1980	Convention on Certain Conventional Weapons
1993	Chemical Weapons Convention
1996	Comprehensive Nuclear Test Ban Treaty
1997-2022	…

* Includes work of precursor bodies: the Ten Nation Committee on Disarmament, the Eighteen National Committee on Disarmament, and the Committee on Disarmament. Years represent the year negotiations ended or the treaty was signed.

Fig. 3 Major agreements of the conference on disarmament

of serious security challenges; when developing advanced systems, such an actor would be less likely to pay a "safety tax."[86]

Finally, will the incentives to take dangerous risks remain high as AI technology improves? With nuclear weapons, this was not the case. Once countries had secure second strike capabilities, the value of a marginal nuclear weapon was relatively small. These diminishing returns enabled the U.S. and Soviet Union to agree to the Strategic Arms Reduction Treaties (START) limiting their numbers of deployed nuclear warheads. There was little incentive to race in developing more. The START treaties are buttressed by the NPT because without the latter, the US and Russia might decide to deploy more warheads to counter more threats. The low incentive to race makes the NPT more valuable and more viable—it provides significant stability and security. If, on the other hand, there were an ever-increasing incentive to deploy more and more warheads, the countries might exhaust themselves in the effort while increasing the chances of conflict.[8788]

In the case of AI, it is unknown whether risks will increase or decrease as the technology progresses. One development scenario starts with developers attempting to align AI systems with the intentions of their creators. While this may prove difficult in the near-term, in the long run, it could be a solved problem with respect to some class of systems. Actors with advanced AI technology may then have little incentive to build systems outside of this class, even if other systems hold out the prospect of a modest increase in capabilities. Non-proliferation would be highly valuable in this world, even sufficient for safety and stability.

Alternatively, if the incentive to develop a more powerful system at the cost of some risk never decreases, then non-proliferation is insufficient for safety and stability.[89] It might still mitigate risks because fewer actors are taking them, but eventually the world will draw the black ball from the urn.[90] Such a world requires other governance options, and it is to those we now turn.

6 Verifiable limits

In this section, we diverge significantly from existing work on AI governance. We think it is important to separate an NPT-based approach from a verification-based one. While the example of the IAEA would seem to suggest they are two sides of the same coin, this example misleads. Fundamentally, the IAEA operates differently for those inside the nuclear club than for those without. The NPT is a great power solution,

[86] Emery-Xu et al. (2023).

[87] On the spiral model in international politics, see Jervis (2017) and Kydd (1997). But, for a counterpoint, see Reiter (1995).

[88] In addition, the dual-use nature of AI incentivizes states to continue development beyond what is necessary for security, blurring the boundary between security-seekers and greedy states. Glaser (1997).

[89] This could occur if advances continue to increase the benefit of a first strike or even merely increase the accuracy of political predictions. Horowitz (2021) and Goldfarb and Lindsay (2021).

[90] Note that other factors influence the value of a non-proliferation regime. If advanced AI facilitates the development of destruction-dominant technologies—those that can cause large-scale destruction against which there is no defense—it is probably better to have those technologies in fewer hands. Non-proliferation regimes that reduce access without fully restricting it can also do more harm than good. A state with lower computing capacity may not pay a "safety tax" in developing advanced systems, for instance.

and the NPT + model we sketch above is intended to *avoid* an intrusive verification regime for leading nations.

6.1 Verifiable limits address more risks

Technology leaders and diplomats have already begun to call for a Verifiable Limits regime.[91] Many fear that proliferation will be almost impossible to check, since AI technologies can spread so easily—the infrastructure necessary to house AI is far more modest than the infrastructure needed to create it. If this is the case, and it will be impossible to contain advanced AI within a club, then an NPT-style regime cannot succeed unless integrated with extensive monitoring and verification (as we consider in this section) or extremely tight control of its physical infrastructure (as we consider in the next section).

Yet, even if an NPT-style regime could successfully limit proliferation, Verifiable Limits might still be more attractive. Many of AI's most severe dangers arise from the process of development itself. Non-proliferation efforts attempt to limit these risks by limiting the number of players. If this would be insufficient to prevent a dangerously misaligned AI, or if risks are not decreasing over time, then limiting development might be the *only* suitable method of governance. Moreover, Verifiable Limits offer a way to privilege expert governance more than NPT + and almost as much as National Standards, since the supervising bodies would necessarily be strictly professional and nonpolitical. Indeed, under International Verification, the need to insulate such a body from the vagaries of international relations could make it among the most nonpolitical international institutions yet devised.

Under the broad heading of Verifiable Limits, we consider two types of regimes: nationally enforced technology caps and International Verification. These models set bounds on the development of certain technologies that might lead harmful forms to emerge. Both involve actors accepting limits that they would not accept if others did not. When it comes to important security technologies where cheating by one actor threatens the security of others, this implies the need for strict verification of compliance with the regime. In the case of AI, these technological limits would almost certainly need to focus on physical and computational infrastructure. Both of these models leave AI development in the hands of states and firms. The models diverge in the extent they internationalize enforcement. A nationally enforced regime capping technology leaves enforcement largely to states: a third party may be necessary to monitor compliance, but it has no authority to terminate or disrupt states' activities. International Verification is bolder: it would give an international body some form of control over AI development by both firms and states, for example by making large AI training runs impossible to execute without its leave.

6.2 There is no example *of major* powers agreeing to analogous verifiable limits

In evaluating the feasibility of Verifiable Limits, it is essential to recognize that history provides no clear precedent for such a regime. In particular, there is no clear example of *major powers* agreeing to either governance model for restricting the development of a powerful military technology for which there is not a military *substitute technology*. Minor powers have certainly agreed to limit various technologies, for instance in the NPT. But these smaller actors rely on the security architecture created by major powers, who can offer greater security guarantees than are afforded by unilateral technological development.[92] Indeed, unilateral arming by smaller actors would tend to lead to a minor power arms race, which would not enhance a minor power's security and would probably undermine it.[93] Major powers, by contrast, have been unwilling to rely on uncertain coalitions of minor powers in lieu of technological development. They have also been unwilling to allow invasive monitoring regimes, which present security risks in themselves.[94]

The record of international covenants might lead a casual observer to think otherwise. The Biological Weapons Convention, for instance, which major powers signed, limits the "development, production and stockpiling" of biological weapons. Yet, this was not a significant military technology without military substitutes: biological weapons' utility was limited because of the potential harm to one's own side, and nuclear weapons offered a clear (and superior) substitute technology. British analysts noted this at the time: "Biological warfare possessed a 'negligible additional deterrent effect next to the megaton bomb' and could not 'make a significant addition to the present deterrent capability of the Western powers.'"[95] Banning biological weapons also *reinforced* the gap between small and major powers because biological weapons were seen as a "poor state's WMD." And tellingly, because the signatories did not believe that a verification regime was possible, the Convention does not include any verification provisions. They signed anyway because violations were not critical security threats.[96] Indeed, the Convention has been flagrantly violated by the Soviet Union and others.

The 1967 Outer Space Treaty, which prohibited deploying nuclear weapons in space, might also appear to be a counterexample. But here, verification was moot because

[91] Kissinger et al., *The Age of AI*, 165.

[92] Narang (2016).

[93] Bas and Coe, "Arms diffusion and war.".

[94] Coe and Vaynman (2020)

[95] Ibid.

[96] Tucker (2002).

space, being transparent, makes verification simple.[97] Moreover, the treaty limited deployment, not development: the distinction is important, because development is both harder to monitor than deployment and harder to reverse if a violation is discovered. Deployment in response to deployment is faster than developing a new research and testing program. Most importantly, the actions the Outer Space Treaty prohibited were not in general thought to be militarily effective: **"**neither weaponization nor war-fighting in space [had] made the transition from fiction to reality."[98] Where they were effective, or might be, the treaty made specific exceptions. Indeed, the superpowers had largely given up on weaponizing space when the treaty was signed.[99] Today, as space weaponization has begun to move out of the realm of science fiction, the major powers have been little deterred by the treaty in weaponizing space.

The Anti-Ballistic Missile (ABM) Treaty repeats the same pattern. It limited the deployment of a technology that "in practice, even without arms control,… would not have changed the strategic balance, although [it] would have led to much wasteful expenditure."[100] Once the technology developed to the point where it began to appear viable, the United States withdrew from the treaty. A full list of international arms control treaties since the mid-nineteenth century is included as Appendix A. It does not include a single precedent for a Verifiable Limits regime that would constrain the major powers.

This is not by chance. States red-team governance proposals from the point of view of how they impact their security interests; even the best-intended proposals will often fall short from this perspective. Nonetheless, Verifiable Limits approaches have been taken seriously at the highest levels of governments. Ronald Reagan and Mikhail Gorbachev, for instance, made proposals along these lines at the 1986 Reykjavík Summit, and similar ideas have been floated for TAI.

7 Red-teaming verifiable limits

A verification treaty for a powerful technology would have a high bar to satisfy the security concerns of major powers. When considering such a proposal for nuclear weapons, the Acheson–Lilienthal Report "concluded unanimously that there is no prospect of security… [in] a system which relies on inspection and similar police-like methods." The report emphasized that their reasons for this conclusion were "*not merely technical,*" that insufficient consideration of the implications of national rivalries were the "fatal defect in the commonly advanced proposals… with a system of inspection," and that this fatal defect "furnished an important clue to us in the development of the plan" for an International Monopoly.[101] Below, we describe in more detail the hurdles any plan for such a regime would need to clear.

Any verification regime must be able to detect prohibited activities in time to stop them from significantly altering the balance of power. This demand poses both a technical and a political problem. At a technical level, monitoring must reliably detect cheating in a project's early stages, for instance, by detecting if an actor has begun to scale up compute beyond acceptable levels.[102] At a political level, states must feel confident that they could disrupt or prevent programs in violation of international law, including by conventional means. Any verification-based proposal must carefully analyze (i) the time to discoverability; (ii) the time to breakout; (iii) the ability of states to disrupt a program within these time frames; (iv) the position of major powers should the verification-based regime unravel, ensuring that these actors are not left worse than before; and (v) the path from the present state to the envisaged regime.

Considering these five factors will influence the design of verification-based regimes. For instance, the sorts of actions that are/are not permitted will influence the expected time to discover violations. Regimes that try to finely parse acceptable from unacceptable behaviors will make timely detection more likely to fail. In the case of AI, trying to parse acceptable uses of computing power will have a harder time uncovering violations than regimes that simply prohibit all uses of computing power in easily identified classes, such as those above a specified number of floating point operations per second (FLOPS).[103] This too was a reason the Acheson-Lilienthal Report recommended internationalization of nuclear technology: the authors believed that attempting to distinguish acceptable from unacceptable uses of nuclear science would be difficult and result in debates that would doom the regime.

Similarly, off-the-path failures may shape how a regime must specify the distribution of computing capabilities, including how vulnerable datacenters are to different actors. States will want to ensure any regime does not reduce their access to these capabilities, especially if disagreements or changes in the technical landscape could cause the regime to fail.

[97] Roger (2010).

[98] Altmann and Scheffran (2003).

[99] Garthoff (1980), Graham and LaVera (2002).

[100] Freedman (2022).

[101] *A report on the international control of atomic energy*. Vol. 2498. US Government Printing Office, 1946. Italics in original.

[102] This is complicated by the fact that current AI systems can be trained on smaller amounts of compute over a longer training period. That is, at a given level of algorithmic progress, there is a Pareto frontier trading off compute and time. See McCandlish et al. (2018).

[103] Though this depends on the ability to run computations more slowly or across distributed systems to avoid detection.

States will also consider how a regime will affect their current standing and influence. They will hesitate to sacrifice technological leads and other advantages over rivals. As well, leaders may be so determined to hang on to advantages that they are unable to accept facts: for example, U.S. President Harry Truman insisted the Soviets would "never" develop their own bomb, despite Robert Oppenheimer's cogent arguments that they soon would.[104] Verifiable Limits might, therefore, be easiest to implement before any one country begins to pull too far ahead or develop significant capabilities it desires to hide. Verifiable Limits tend to level the AI field among all nations. The rougher that field to begin with, the more difficult leveling it will be. If China or the United States enjoys a significant lead, it will be hard to find a *quid pro quo* sufficient to induce it to sacrifice that lead for the sake of international stability. As we discussed when introducing the taxonomy, some degree of uncertainty may be necessary to get governance off the ground.

Finally, when states cede authority to an international body, the governance of that body is a particular point of concern. In practice, institutions often mirror the balance of international power at their creation. While this practice makes it easier to create institutions, as time passes they cease to reflect geopolitical realities. As power and influence shift, the regime can become increasingly outmoded and, as a result, increasingly contested.

An apparent feature of Verifiable Limits is the way it balances between a variety of actors. But this feature may be a bug. If transformative AI turns out to enjoy significant economies of scale, then the market (or the balance of power) may tend to reduce the dangers of an AI race over time: as some actors draw farther ahead and consolidate their dominant positions, it would become less and less possible for other actors to close the gap. In this scenario, actors face steadily decreasing incentives to cut corners, and so accidental risks would be declining with time. Likewise, if TAI turns out to be destruction dominant, such economies of scale may also promote peace through deterrence.[105] Imposing artificial caps on AI development would undermine this salutary dynamic: these caps could inflame AI races by keeping competitors near parity. In an extreme case, a Verifiable Limits regime might wind up *maximizing* the number of competitors on the cusp of transformative advances in AI, and so maximize the probability some of them cheat or eliminate safety precautions. Finally, such caps might also impede or even foreclose other governance options like an International Monopoly, especially if such a regime would need to evolve at least partially on its own. Indeed, without caps on its development, a monopoly may be the natural endpoint for AI due to economies of scale, and such a monopoly might actually be beneficial for world politics.

8 Technical solutions

Some of these issues might be resolved through technical solutions. Cryptographic techniques might ease problems of monitoring and verification.[106] They might facilitate verification in such a way that additional information about state activities does not become public. Alternatively, a technical equivalent of the "two person rule" from nuclear launch protocol might be available for the control of datacenters by actors separated in space and time. If certain training runs might be dangerous for one or all actors, such a technical solution could imply that training runs would only occur if an entire set of actors agreed.

Technical approaches to facilitate agreement also face challenges that go beyond technical feasibility, however. For instance, they would need to be credible to a range of adversarial actors. This implies a need for transparency: an adversary must be confident the technology does what is promised. But technological transparency also facilitates malicious interference. Whether these circles can be squared remains an essential area for research.

A final point here is worth noting. At the moment, firms dominate AI research. The political obstacles to Verifiable Limits are more manageable when governing AI research by firms than by states. If firms remain in the driver's seat, states might *prefer* to outsource monitoring to a third party accountable to other actors: doing so allows them to reassure other actors that their firms are obeying the rules, both because third parties can provide more credible information[107] and because monitoring can be divorced from domestic politics. In such a world, the technical apparatus necessary to sustain an ambitious International Verification regime might be within reach. Nonetheless, we do not think it likely that the soft touch of the US and other governments will continue toward AI. It would be dangerous, perhaps even naive, to build a regime expecting that the primary actors will not include states.

In short, Verifiable Limits face three challenges. First, the surveillance apparatus necessary to reliably detect cheating may be impossible to construct; in the case of AI, detection may be too unreliable when the physical infrastructure and professional knowhow are widely dispersed, including on ordinary devices. Second, even if it is possible, at least one great power might refuse to submit to it, in which case global cooperation would likely fail. Finally, Verifiable Limits might impede, or even prevent, a monopoly from emerging.

[104] Monk.

[105] Deterrence may reduce the risk of great power war but may not lower the rate of low level conflict. Lee et al. (2023).

[106] Shavit (2023).

[107] Keohane (1984).

To see why a monopoly might be attractive, rather than a fate to avoid, we now consider it as the final possibility for governing a transformative technology.

9 International monopoly

Verifiable Limits will fail if the necessary inspection regime is too invasive or too unreliable. A monopoly removes this difficulty by consolidating the epistemic or physical resources necessary to develop transformative technology in a single international body. Ideally, this body would be governed by international covenants, and all nations (subject to certain rules) would benefit from its activities. While we locate this proposal at the proliferation stage, since it describes how many actors possess a technology, it should be noted that the existence of a monopoly could make limits on development much easier to implement, especially for a technology like AI.

Consolidating a transformative technology within an International Monopoly is not a new idea. The Acheson–Lilienthal Commission concluded unanimously that, despite the apparent risks of ceding so much power to a new international organization, a monopoly was the only form of international governance that was equal to the nuclear task. Their proposal grew out of a variety of suggestions, some more feasible than others, for governing the atomic age. These suggestions began with Niels Bohr.

Bohr imagined an "open world," one where a single community of scientists spanned the entire globe. To advance this vision, he urged Franklin Roosevelt to share nuclear secrets with the Soviet Union, and he converted many of his fellow scientists to his dream of postwar order. In his vision, "the United States and the United Kingdom would 'trade' their atomic knowledge for an open world."[108] This world would escape the arms races and standoffs that became endemic in the Cold War: because there would be no secrets, there would be no suspicion; and because there would be no nuclear arms, there would be no arms race.

Echoes of Bohr's vision can be seen in the heady open-source days of AI's infancy. Nonetheless, even if Bohr's dream could have governed atomic mysteries, such an open world might be undesirable for other transformative technologies like advanced AI. (If AI uncovers destruction-dominant technologies, a single sociopath could wreak such damage as to make some control of the technology a matter of life-and-death.) Like the Acheson Proposal, an International Monopoly on TAI seeks to retain something of Bohr's idealism while accommodating realities of human nature and international politics.

9.1 A monopoly might mitigate risk with less-invasive monitoring

By monopolizing talent or resources in an international body, this regime reduces the level of verification necessary to sustain cooperation. For instance, because technology caps leave computational infrastructure in the hands of states or firms, they require some way to monitor how those actors are using it. Both a nationally enforced technology cap and an artificially created International Monopoly might require tracking chips to ensure that no actors are developing unsanctioned capabilities. With less control over the chip supply chain, however, technology caps may require more invasive surveillance of state activities. Unlike technology caps, an International Monopoly would not require close supervision of state training runs to determine whether the state is developing a sanctioned or unsanctioned system. An International Monopoly is similar to proposals for centralizing compute resources within states, only on a global scale. It also resembles proposals for creating large national compute resources (like the National Artificial Intelligence Research Resource (NAIRR)), except that it excludes other actors from possessing such capabilities themselves.[109] As in NAIRR, significant aspects of AI development might remain decentralized in an International Monopoly, including the code itself and data analysis. The key is only to make it impossible to create TAI without involving the monopoly.

Creating this sort of monopoly artificially likely requires establishing control over some essential inputs. The Acheson–Lilienthal Report proposed to do this by establishing control over all aspects of uranium and thorium, including its mines. An International AI Monopoly might establish control over relevant advanced chip supply chains and monitor their distribution.[110] It might also take further measures, like maintaining secrecy with respect to training techniques, including RL environments and data curation. In the case of natural monopolies that derive from economies of scale, these measures might be unnecessary.

If we believe that AI's economies of scale will continue to grow, then rogue actors might be deterred not through the threat of punishment, but through the sheer magnitude of the international community's headstart—creating a self-enforcing equilibrium. Indeed, if major actors like the US

[108] Bird and Sherwin (2005).

[109] National Artificial Intelligence Research Resource Task Force (2023).

[110] A key caveat to such a proposal would be continued algorithmic progress, which in the field of computer vision halves the compute requirement every 6 months for training a model. Despite this, the computational burden of state-of-the-art models continues to increase. In the limit, whether consumer electronics will ever be able to train a transformative AI model depends on as-yet unknown physical limits to transistor density and theoretical limits to algorithms' computational complexity. Erdil and Besiroglu (2022).

and EU pooled their resources, it might be possible for them to pull so far ahead that even other great powers could not hope to catch up. Certainly, the invasiveness necessary to sustain such a regime would be much less than under any verification-based scheme. It might still, though, be unacceptable to the great powers, depending on how easily the technology could be disseminated and how dangerous rogue actors might be.[111]

The purported advantages of an International Monopoly are many, and we do not attempt to weigh all of the tradeoffs. Yet, a few of the purported advantages are worth noting. Like International Verification, it privileges expert governance, and this expertise is relatively insulated from political pressure. Like the Acheson–Lilienthal Plan, it averts dangerous arms races, and it might reduce mutual suspicion, which can fester even when all actors are well-intentioned. It also asks states to relinquish a valuable technology by promising to keep the technology in hands that, proponents hope, could not turn it to political advantage: just as an international scientific agency would lack the means to deliver a warhead against a state, an International Monopoly might lack the kinetic infrastructure necessary to set itself up as a global hegemon.

9.2 Security red-teaming an international monopoly

Nonetheless, the same security factors that confront Verifiable Limits approaches also confront International Monopolies. Before ceding some portion of responsibility for their security to an international body, states would ask, for one, whether cheating by a rival would be possible.

Like the other approaches to supranational governance, an International Monopoly has substantial potential downsides. Most obviously, its inspiration, the Acheson–Lilienthal Proposal, failed, and it is far from certain that there is an updated version that could succeed. Its potential concentration of power poses clear risks. Steps would need to be taken, probably of a technical nature, to prevent its misuse. If the international organization does not usurp power itself, a state or group of states could wield awesome power by politically capturing the body and excluding others from its benefits. States of the Global South would certainly hesitate to trust developed nations with such power. For such options to be contemplated, the common interest in reducing competition over advanced AI would need to be acute. Or, such a monopoly would need to evolve on its own. This may be one advantage that an AI monopoly has over the Acheson–Lilienthal Plan: nuclear technology does not exhibit strong tendencies toward market concentration, but advanced AI might.

10 Technical solutions

As with Verifiable Limits, there may be technical solutions to many of these concerns. Some researchers have begun exploring whether AI training runs could be made to require "keys," so that major powers would have to consent before any run could execute. This suggestion mimics nuclear weapons systems, which often require two human beings to activate. It also seems more plausible that countries like the US and China would be more willing to give the other a key over a third party than over its own network. In our view, technical solutions of this kind hold great promise. If they prove cryptographically feasible, they will mitigate fears about losing control of the technology and of the institution governing it. In this case, monopolizing technology inputs may reduce fears of some party making surreptitious advances, and thus reduce incentives to race.

We should also not overlook potentially low-tech solutions to these high-tech problems. If an International Monopoly maintains its physical infrastructure in a vulnerable location, then any major power could retain an effective "killswitch" over its development activities. Credibly threatening a conventional strike might suffice to reset the balance of power.[112]

Something like this last, low-tech suggestion currently obtains with respect to the future flow of semiconductors. As the supply chain disruptions of the past few years have made very clear, a shockingly large proportion of advanced semiconductors are produced in a single country, Taiwan, and this country is highly vulnerable to the world's leading powers.

But Taiwan's example should remind us why some degree of *design* is vital for AI governance. That semiconductor production was destined to be concentrated with a few firms, operating in the same east Asian country, was perhaps an inevitable outworking of market evolution. But the precise real estate where this concentration occurred might easily

[111] For instance, if even modest AI enables rogue actors to develop powerful bioweapons, as is likely to be the case, then it would not suffice for a monopoly simply to be able to maintain itself at the cutting edge of TAI. Rather, some form of global verification would be necessary. See for instance Dunlap and Pauwels (2017).

[112] While such capabilities are most feasible in an International Monopoly, they can in fact also be exploited in other governance models including NPT + and Verifiable Limits. It is more difficult, politically, to achieve the same degree of vulnerability and transparency, but there are precedents. The ABM Treaty, for instance, contains a provision that "Each Party undertakes not to use deliberate concealment measures which impede verification by national technical means of compliance with the provisions of this Treaty." Similar language might be used, for instance, to ensure the visibility and vulnerability of large data centers.

have been elsewhere. If TAI is left to evolve on its own, it may well be concentrated in very few hands. There is no guarantee those hands will be benign, nor that they will be free of other fraught political questions (as Taiwan's example demonstrates). Indeed, if TAI is powerful enough to grant its creator a decisive strategic advantage, as nuclear weapons arguably did briefly for the United States, it is quite possible such an actor would use that advantage to rewrite the rules of international order. It seems better, if social and technical challenges can be overcome, to implement a broader, consent-based governance regime before that can happen.

Before concluding, it is worth considering whether milder versions of internationalization might be feasible. Some forms of supranational governance would threaten state sovereignty and the balance of power much less than an International Monopoly. Incentives for adopting international standards into domestic regulations, which exist in many industries, would reduce the race-to-the-bottom concerns of a National Standards approach. International institutions can govern civilian development and use when domestic governments support international regimes through tying market access to standards compliance. This occurs in industries like civil aviation, maritime, and areas of finance.[113] Proposals of this sort can help to bridge the divide between the AI "haves" and "have-nots." These are important benefits, and we believe such approaches hold great promise for ameliorating risks in civilian AI, if not eliminating them. Unfortunately, we do not believe such approaches are sufficient to address the risks that we describe above of competition in the development of military technologies. Without restricting access to military TAI (as in NPT+or monopoly) or monitoring its development (Verifiable Limits), international governance would not address some of the most significant risks surrounding the technology. It would not stop AI races in the military domain. International governance of civilian technologies may be a useful tool for protecting privacy, preventing human rights abuses, ensuring the safety of advanced civilian systems, and other governance objectives, but it likely cannot form the bedrock of an effective regime for governing the development and use of AI for security, any more than such a regime could hope to govern nuclear arms.

11 Paths forward for governing TAI

Currently, subnational groups are spearheading AI governance, and socialization remains their preferred tool. While social norms have often proved insufficient to govern emerging technologies, we have reasons to believe that they will be more effective in the years ahead. Nonetheless, there is a high likelihood that they will not suffice to govern AI on their own: the technology has too many power-political implications, and the rewards are too great, for professional and social sanctions to constrain its development. Moreover, as AI has progressed, state actors have paid increasing attention to this once-ignored domain; it is unclear how much longer nonstate actors will remain in the driver's seat. The speed with which AI is progressing leaves norms little time to evolve.

If substate actors will not be the primary *who* of AI governance when it comes to security, it would seem natural to turn to traditionally realist or liberal internationalist models instead. And indeed, all four of the models we see as potentially viable for security governance among states (NPT+, Verifiable Limits, International Monopoly, International Hegemony) fit comfortably within these traditions. The old paradigms, though, may be misleading. In both the realist and liberal traditions, cooperation is often built on reciprocity.[114] Reciprocity, in turn, only succeeds as part of a repeated game: deviations are prevented by the threat of future retaliation, as in the classic tit-for-tat solution to the Prisoner's Dilemma. Yet, new technologies sometimes present an unusual challenge to these familiar patterns. With TAI, as with nuclear weapons in a world without them, it is possible that a single deviation from the regime may be sufficiently powerful to make future retaliation impossible.

Moreover, any regime must be stable in both the short and long run. It must be *dynamically* stable, and so capable of adapting as TAI reconfigures the sociopolitical terrain. For instance, NPT+is easier to implement because it does not disrupt the balance of power; but will it remain viable when this balance itself begins shifting? Similarly, the computational power necessary to execute a given AI training run decreases significantly every year; can a verification-based regime survive, when it becomes easier and easier to evade? These questions suggest that both socialization and nonstate actors will play a vital role in the long-term success of any AI regime. Despite the diminishing barriers to entry, few countries have developed nuclear weapons, and their restraint seems to stem in significant part from international norms, not cynical calculation.[115] In the same way, we suspect that norms and nonstate actors, although they will be unable to limit the development of TAI in its early days, *will* be able, in the decades ahead, to stabilize, entrench, and refine whatever regime does emerge.

Among multilateral AI governance regimes, we see NPT+as the most feasible. It has the strongest historical precedent. Because many actors are in the "club" already, it minimizes the incentives to break out of the regime. It would seem to avoid invasive monitoring, and it would

[113] Trager et al. (2023).

[114] For classic treatments, see Schelling (1966) and Axelrod (1984).

[115] Rublee (2009) and Kemp (2014).

preserve (or even calcify) the current balance of power, making it attractive to leading states. Moreover, it could arise either by design or by gradual cooperative evolution, and it requires less initial buy-in than a verification regime or internationalization. It also aligns comfortably with the economies of scale that TAI may exhibit. Nonetheless, NPT + assumes that risks plateau because the actors in the technology club do not have continuing incentives to "cut corners" to improve their relative power. NPT + is insufficient in a world where TAI poses large risks due to competition among members of the club and such a world may be plausible.

Ultimately, we cannot yet know which of these four models will be most effective. We are, after all, still in the first half of the Collingridge Dilemma. Nevertheless, there are actions that could be taken today that would yield a high return across many scenarios. One is to investigate how existing institutions might be deployed in the service of the governance models we discuss above. In the current international environment, it may be impossible to achieve grand, broadly inclusive treaties to govern AI. Another approach is to investigate how broad governance regimes might grow out of bilateral and club models. Bilateral engagement between the U.S. and Soviet Union eventually produced the NPT, which currently has 190 signatories. The Financial Action Task Force began in the G7 and now has 40 member states and many other cooperating parties.

Another set of near-term actions includes investigating technical mechanisms that facilitate governance and increase governance options.[116] Some form of verification will probably figure in any regime, for instance, and this verification will probably involve governing the compute to develop and deploy models that influence state security. Many potential agreements could require methods of monitoring the distribution and activities of advanced chips, i.e., those capable of being networked together to produce TAI. This might require tracking hardware built into individual chips and know-thy-customer accounting regimes throughout the semiconductor supply chain. Such measures would enable many governance options because they make breakout more difficult, and reduce the necessary invasiveness of whatever verification actors might later wish to implement. In some scenarios, such capacities could facilitate trust between cooperating actors. Still, whether verification can become the foundation of a governance regime rather than just its accessory will depend on verification technologies whose feasibility we do not yet know; it will also depend on how much scrutiny political leaders and great powers are willing to countenance.

There is, as we explained above, no precedent for such governance regimes. Nonetheless, while there is no precedent for invasive monitoring of the great powers, there is also no precedent for "intelligent" technology. If the technology evolves quickly, we are about to live through unprecedented times; it should not surprise us if unprecedented political structures emerge—though we should not blithely assume they will.

Appendix A: Arms Control Treaties Since 1868

Treaty name	Year signed/ entered into force?	Verification regime
TPNW	2017	Potential inspections by authority not yet designated
New START	2010	National Technical Means
CCM	2008	No provision
NWFZ in Central Asia	2006	No provision
CCW Protocol V	2005	No provision
SORT/Moscow Treaty	2002	No provision
Ottawa Convention	1997	Missions approved by majority
Reduction of Forces in Border Areas	1997	No provision
CTBT	1996	state inspections
African NWFZ	1996	IAEA inspection
Southeast Asia NWFZ	1995	IAEA inspection
CCW Protocol IV – BLWs	1995	No provision
CWC	1993	intrusive verification regime
START II	1993	No provision
OST	1992	No provision
Mongolia NWFZ	1992	IAEA inspections
START I	1991	National Technical Means
CFE	1990	National Technical Means
U.S.-U.S.S.R. CW Destruction	1990	Inspections
INF	1987	National Technical Means + inspections
South Pacific NWFZ	1985	IAEA provisions
CCW Protocol III	1981	No provision
CCW Protocol II		No provision
CCW Protocol I		No provision
SALT II	1979	NTM
SALT I	1972	National Technical Means
BWC	1972	National Technical Means

[116] Reuel et al. (2024).

Treaty name	Year signed/entered into force?	Verification regime
ABM Treaty	1972	National Technical Means
Seabed Arms Control Treaty	1971	National Technical Means
NPT	1968	IAEA inspections
Latin American NWFZ	1967	IAEA inspections
Outer Space Treaty	1967	National Technical Means
The Antarctic Treaty	1959	National Technical Means
The Geneva Protocol	1925	No provision
Washington Naval Treaty	1922	No provision
Convention … Laying of Submarine Mines	1907	No provision
Convention … Bombardment by Naval Forces in Time of War	1907	No provision
Declaration Prohibiting the Discharge of Projectiles and Explosives from Balloons	1907	No provision
Declaration … Prohibition of the Discharge of Projectiles and Explosives from Balloons …	1889	No provision
Declaration …Prohibition of the Use of Projectiles with the Sole Object to Spread Asphyxiating Poisonous Gases	1889	No provision
Declaration …Prohibition of the Use of Bullets which can Easily Expand	1889	No provision
St Petersburg Declaration	1868	No provision

Declarations

Conflict of interest The authors have no competing interests to declare.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Abbott KW (2012) The transnational regime complex for climate change. Environ Plann C Polit Space 30(4):571–590

Ahmed N, Wahed M (2020) The de-democratization of AI: deep learning and the compute divide in artificial intelligence research. arXiv:2010.15581 [cs.CY]

Allen GC (2022) Choking off China's access to the future of AI. Center for Strategic & International Studies. https://www.csis.org/analysis/choking-chinas-access-future-ai

Altman S (2022) Average is probably single-digits cents per chat; trying to figure out more precisely and also how we can optimize it (Twitter, 2022). https://twitter.com/sama/status/1599671496636780546?lang=en

Altmann J, Scheffran J (2003) New rules in outer space: options and scenarios. Secur Dial 34(1):109–116

Armstrong S, Bostrom N, Shulman C (2016) Racing to the precipice: a model of artificial intelligence development. AI Soc 31:201–206

Axelrod R (1984) The evolution of cooperation. Basic Books, New York

Bas MA, Coe AJ (2012) Arms diffusion and war. J Conflict Resolut 56(4):651–674

Ben Garfinkel B, Dafoe A (2019) How does the offense-defense balance scale? J Strat Stud 42(6):736–763

Beraja M, Kao A, Yang DY, Yuchtman N (2023) AI-tocracy. Q J Econ. https://doi.org/10.1093/qje/qjad012

Bertuzzi L (2023) EU's AI Act negotiations hit the brakes over foundation models. Euractiv Nov 15:2023

Besiroglu T, Emery-Xu N, Thompson N (2024) Economic impacts of AI-augmented R&D. Res Policy 53(7)

Biddle S (2004) Military power: explaining victory and defeat in modern battle. Princeton University Press, Princeton

Bird K, Sherwin MJ (2005) American prometheus: the triumph and tragedy of Robert J. Oppenheimer. Vintage Books, New York, p 275

Brundage M et al (2018) The malicious use of artificial intelligence. arXiv:1802.07228 [cs.AI]

Buchanan B (2020) The AI triad and what it means for national security strategy. Center for Security and Emerging Technology. https://doi.org/10.51593/20200021

Bull H (1973) The Moscow agreements and strategic arms limitation. Australian National University Press

Cebul MD, Dafoe A, Monteiro NP (2021) Coercion and the credibility of assurances. J Polit 83(3)

Chan K (2022) EU court largely upholds $4B Google Android antitrust fine. AP News

Coe AJ, Vaynman J (2015) Collusion and the nuclear nonproliferation regime. J Polit 77(4):889–1175

Coe AK, Vaynman J (2020) Why arms control is so rare. Am Polit Sci Rev 114(2):342–355

Cohen J (2018) After last week's shock, scientists scramble to prevent more gene-edited babies. Science. https://www.science.org/content/article/after-last-weeks-shock-scientists-scramble-prevent-more-gene-edited-babies

Collingridge D (1980) The social control of technology. St Martin's Press, New York

Crafts N (2021) Artificial intelligence as a general-purpose technology: an historical perspective. Oxf Rev Econ Policy 37(3)

Crawford TW, Vu KX (2021) Arms control as wedge strategy: how arms limitation deals divide alliances. Int Secur 46(2):91–129

Dafoe A (2018) AI Governance: a research agenda. Future of Humanity Institute, Oxford

Dunlap G, Pauwels E (2017) The intelligent and connected bio-labs of the future: promise and peril in the fourth industrial revolution. The Wilson Center, Washington, D.C.

Eckersley R (2012) Moving forward in the climate negotiations: multilateralism or minilateralism? Glob Environ Polit 12(2):24–42

Erdil E, Besiroglu T (2022) Algorithmic progress in computer vision. arXiv:2212.05153

Emery-Xu N et al (2023) Uncertainty, information, and risk in international technology races. J Conflict Resol 0(0):1–29

Farrell H, Newman A (2016) The new interdependence approach: theoretical development and empirical demonstration. Rev Int Political Econ 23(5):713–736

Fredrickson DS (1991) Asilomar and recombinant DNA: the end of the beginning. In: Hanna KE (ed) Biomedical politics. National Academy Press, Washington, p 284

Freedman L (2022) SALT 50 years on: strategic theory and arms control. Survival 64(2):56

Garthoff RL (1980) Banning the bomb in outer space. Int Secur 5(3):25–40

Gavin FJ (2015) Strategies of inhibition: US grand strategy, the nuclear revolution, and nonproliferation. Int Secur 40(1):9–46

Geist E, Lohn AJ (2018) How might artificial intelligence affect the risk of nuclear war? RAND Corporation, Santa Monica. https://www.rand.org/pubs/perspectives/PE296.html

Gilli A, Gilli M (2018/19) Why China has not caught up yet: military-technological superiority and the limits of imitation, reverse engineering, and cyber espionage. Int Secur 43(3):141–189. https://doi.org/10.1162/isec_a_00337

Glaser CL (1997) The security dilemma revisited. World Polit 50(1):171–201

Glaser CL, Kaufmann C (1998) What is the offense-defense balance and can we measure it? Int Secur 22(4):44–82

Goldfarb A, Lindsay JR (2021) Prediction and judgment: why artificial intelligence increases the importance of humans in war. Int Secur 46(3):7–50

Grace K, Salvatier J, Dafoe A, Zhang B, Evans O (2018) Viewpoint: when will AI exceed human performance? Evidence from AI experts. J Artif Intell Res. https://doi.org/10.1613/jair.1.11222

Graham T, LaVera DJ (2002) Cornerstones of security arms control treaties in the nuclear era. University of Washington Press, Seattle

Grant N (2023) Google calls in help from Larry Page and Sergey Brin for A.I. Fight, New York Times

Grissom A (2006) The future of military innovation studies. J Strateg Stud 29(5):905–934. https://doi.org/10.1080/01402390600901067

Gruetzemacher R, Whittlestone J (2019) Defining and unpacking transformative AI (unpublished manuscript)

Gruetzemacher R, Whittlestone J (2022) The transformative potential of artificial intelligence. Futures 135

Hale T, Held D, Young K (2013) Gridlock: why global cooperation is failing when we need it most

Hamburg MA, Jaime Yassif R, Charo A, Severance H (2022) Taking action to safeguard bioscience and protect against future global biological risks. Sci Diplom. https://doi.org/10.1126/scidip.ade6829

Hoffman J et al (2022) Training compute-optimal large language models. https://arxiv.org/abs/2203.15556 [cs.CL]

Hopf T (1991) Polarity, the offense-defense balance, and war. Am Polit Sci Rev 85(2):475–493. https://doi.org/10.2307/1963170

Horowitz MC (2010) The diffusion of military power: causes and consequences for international politics. Princeton University Press, Princeton

Horowitz MC (2018) Artificial intelligence, international competition, and the balance of power. Texas Natl Secur Rev 1(3)

Horowitz MC (2020) Do emerging military technologies matter for international politics? Annu Rev Polit Sci 23:385–400. https://doi.org/10.1146/annurev-polisci-050718-032725

Horowitz MC (2021) When speed kills: lethal autonomous weapon systems, deterrence and stability. In: Sechser T, Narang N, Talmadge C (eds) Emerging technologies and international stability. Routledge, London, pp 144–168

Howard M (2009) War in European history. Oxford University Press, Oxford

Jervis R (1978a) Cooperation under the security dilemma. World Polit 30(2):167–214. https://doi.org/10.2307/2009958

Jervis R (1978b) Cooperation under the security dilemma. World Polit 30(2):167–214

Jervis R (2017) Perception and misperception in international politics. Princeton University Press, Princeton

John Ikenberry G, Kupchan CA (1990) Socialization and hegemonic power. Int Organ 44(3):283–315

Abbot KW, Faude B (2021) Choosing low-cost institutions in global governance. Int Theory 13(3):397–426

Keohane RO (1984) After hegemony. Princeton University Press, Princeton

Khan SM, Mann A (2020) AI chips: what they are and why they matter. Center for Security and Emerging Technology

Kissinger HA, Schmidt E, Huttenlocher D (2021) The age of AI and our human future. Little, Brown and Company, New York, p 166

Krimsky S (1982) Genetic alchemy: the social history of the recombinant DNA controversy. MIT Press, Cambridge

Kroenig M (2014) Force or friendship? Explaining great power nonproliferation policy. Secur Stud 23(1):1–32

Kydd A (1997) Game theory and the spiral model. World Polit

Lee KS, Kim JD, Jin H, Fuhrmann M (2023) Nuclear weapons and low-level military conflict. Int Stud Q 66(5)

Logan DC (2022) The nuclear balance is what states make of it. Int Secur 46(4):172–215

Lohn AJ, Jackson KA (2022) Will AI make cyber swords or shields? Center for Security and Emerging Technology. https://doi.org/10.51593/2022CA002

Mayer B, Zahar A (eds) (2021) Debating climate law. Cambridge University Press, Cambridge

McCandlish S, Kaplan J, Amodei D, OpenAI Dota Team (2018) An empirical model of large-batch training. arXiv:1812.06162

Meirowitz A, Sartori AE (2008) Strategic uncertainty as a cause of war. Q J Polit Sci 3(4):327–352

Mirsky Y, Lee W (2021) The creation and detection of deepfakes: a survey. ACM Comput Surv 54(1):1–41. https://doi.org/10.1145/3425780

Mitchell C (2022) The Power of Delay: Banking System Sturcture and Implementation of the Basel Accords. Bus Politics 24(1):1–17

Morton F (1989) Thunder at Twilight. Scribners, New York, p 320

Narang V (2016) Strategies of nuclear proliferation: How states pursue the bomb. Int Secur 41(3):110–150

National Artificial Intelligence Research Resource Task Force (2023) Strengthening and democratizing the U.S. artificial intelligence innovation ecosystem: an implementation plan for a National Artificial Intelligence Research Resource. National Science Foundation, Washington, D.C. https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf

Ord T (2020) The precipice: existential risk and the future of humanity. Hachette Books

Partzsch L (2020) Alternatives to multilateralism: new forms of social and environmental governance. MIT Press, Cambridge

Prunkl CEA, Ashurst C, Anderljung M, Webb H, Leike J, Dafoe A (2021) Institutionalizing ethics in AI through broader impact requirements. Nat Mach Intell 3(2):104–110

Reiter D (1995) Exploding the powder keg myth: preemptive wars almost never happen. Int Secur 20(2):5–34

Rublee MR (2009) Nonproliferation norms: why states choose nuclear restraint. University of Georgia Press, Atlanta

Reuel A, Bucknall B et al (2024) Open problems in technical AI governance. arXiv:2407.14981

Rodgers M (1977) Biohazard. Random House Inc., p 52

Roger HG (2010) Space and verification, volume I: policy implications (November 9, 2010). Eisenhower Center for Space and Defense Studies. https://swfound.org/media/37101/space%20and%20verification%20vol%201%20-%20policy%20implications.pdf; “The People's Republic of China and the Russian Federation working paper: verification aspects of PAROS”, p. 3, para. 6, https://www.un.org/disarmament/wp-content/uploads/2018/04/A-CN.10-2018-WG.2-CRP.2-E2.pdf

Russell S (2019) Human compatible: artificial intelligence and the problem of control. Penguin, London

Sagan SD (1994) The perils of proliferation: organization theory, deterrence theory, and the spread of nuclear weapons. Int Secur 18(4):66–107

Scao TL et al (2022) BLOOM: a 176B-parameter open-access multilingual language model. https://arxiv.org/abs/2211.05100 [cs.CL]

Schelling TC (1966) Arms and influence. Yale University Press, New Haven

Schelling T, Halperin MH (1961) Strategy and arms control. The Twentieth Century Fund, New York

Schiffer Z (2021) Google fires second AI ethics researcher following internal investigation. The Verge February 19, 2021, https://www.theverge.com/2021/2/19/22292011/google-second-ethical-ai-researcher-fired

Schiffer Z, Newton C (2023) Microsoft lays off team that taught employees how to make AI tools responsibly. The Verge. https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs

Scott Kemp R (2014) The nonproliferation emperor has no clothes: the gas centrifuge, supply-side controls, and the future of nuclear proliferation. Int Secur 38(4):39–78

Shane S, Wakabayashi D (2018) ‘The Business of War’: Google employees protest work for the pentagon. New York Times

Shavit Y (2023) What does it take to catch a Chinchilla? Verifying rules on large-scale neural network training via compute monitoring. arXiv:2303.11341

Shevlane T (2022) Structured access: an emerging paradigm for safe AI deployment. In: Bullock JB et al (eds) The Oxford handbook of AI governance. Oxford. https://doi.org/10.1093/oxfordhb/9780197579329.013.39

Slaughter A-M (1997) The real new world order. Foreign Aff

Snidal D (1991) International cooperation among relative gains maximizers. Int Stud Q 35(4):387–402

Solingen E (2009) Nuclear logics. Princeton University Press, Princeton

Solow R (1987) We’d better watch out. The New York Times, p 36. http://www.standupeconomist.com/pdf/misc/solow-computer-productivity.pdf

Stafford E, Trager RF (2022) The IAEA solution: knowledge sharing to prevent dangerous technology races, working paper

Stern J (2002) Dreaded risks and the control of biological weapons. Int Secur 27(3):89–123

Stern J (2003) Dreaded risks and the control of biological weapons. Int Secur 27(3):89–123

Thompson NC et al (2022) The computational limits of deep learning. https://arxiv.org/abs/2007.05558 [cs.LG]

Trager R, Stafford E, Emery-Xu N, Dafoe A (2022) Safety in technology competitions, working paper

Trager R et al (2023) International governance of civilian AI: a jurisdictional certification approach. arXiv:2308.15514

Tucker JB (2002) In the shadow of anthrax: strengthening the biological disarmament regime. Nonprolif Rev 9(1):112–121. https://doi.org/10.1080/10736700208436877

Urbina F, Lentzos F, Invernizzi C, Ekins S (2022) Dual use of artificial-intelligence-powered drug discovery. Nat Mach Intell 4(3):189–191

Webb A (2019) The big nine: how the tech titans and their thinking machines could warp humanity. Public Affairs, New York

Weidinger L et al (2021) Ethical and social risks of harm from language models. https://arxiv.org/abs/2112.04359. [cs.CL]

Zhang B, Dreksler N, Anderljung M, Kahn L, Giattino C, Dafoe A, Horowitz MC (2022) Forecasting AI progress: evidence from a survey of machine learning researchers. arXiv:2206.04132v1 [cs.CY]

Zwetsloot R, Dafoe A (2019) Thinking about risks from AI: accidents, misuse and structure. Lawfare

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.