Expert Insights
PERSPECTIVE ON A TIMELY POLICY ISSUE

BENJAMIN BOUDREAUX, GREGORY SMITH, EDWARD GEIST, LEAH DION

Insights from Nuclear History for AI Governance

May 2025

For more information on this publication, visit **www.rand.org/t/PEA3652-1**.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.

About This Paper

There have been multiple proposals for the international governance of artificial intelligence (AI) that draw from the existing nuclear governance regimes. In this paper, we analyze lessons from the history of nuclear stability and draw analogies to building international governance of AI. We analyze two major episodes in nuclear governance, the failure of the Baruch Plan and the success of the Non-Proliferation Treaty, to understand what factors led to the failure or success of these governance initiatives. We then identify the challenges that proposals for global AI governance face that might make building a regime similar to the nuclear nonproliferation one difficult. This paper is intended for those interested in potential models for global governance of AI that draw on past global governance efforts, such as nuclear nonproliferation.

Technology and Security Policy Center

RAND Global and Emerging Risks is a division of RAND that delivers rigorous and objective public policy research on the most consequential challenges to civilization and global security. This work was undertaken by the division's Technology and Security Policy Center, which explores how high-consequence, dual-use technologies change the global competition and threat environment, then develops policy and technology options to advance the security of the United States, its allies and partners, and the world. For more information, contact tasp@rand.org.

Funding

This effort was, at its inception, independently initiated and conducted within the Technology and Security Policy Center using income from operations and gifts from philanthropic supporters, which have been made or recommended by DALHAP Investments Ltd., Effektiv Spenden, Ergo Impact, Founders Pledge, Charlottes och Fredriks Stiftelse, Good Ventures, Jaan Tallinn, Longview, Open Philanthropy, and Waking Up Foundation. This effort was published through funding provided by DALHAP Investments Ltd., as recommended by Ergo Impact and Fathom. A complete list of donors and funders is available at www.rand.org/TASP. RAND donors and grantors have no influence over research findings or recommendations.

Acknowledgments

We thank the leadership of the RAND Technology and Security Policy Center, Jeff Alstott and Emma Westerman, for their guidance on this publication. We also thank Brodi Kotila for her guidance and oversight of this publication and our reviewers, Matthew Chessen and Robert Trager.

Contents

Insights from Nuclear History for AI Governance

Developments and applications of artificial intelligence (AI) technologies have motivated calls for domestic and international regulatory regimes to address risks. These risks include, in the most extreme prognosis, possible catastrophic harm to humanity. Proponents cite various reasons for establishing new international governance mechanisms. Sam Altman and others at OpenAI have expressed fears of uncontrollable superintelligent systems that might require a new international authority.[1] Others argue that different types of AI risks necessitate international governance. For instance, United Nations (UN) Secretary-General António Guterres has stressed how generative AI "could be a defining moment for disinformation and hate speech" and could enable new levels of authoritarian surveillance; he notes that "without action to address these risks, we are derelict in our responsibilities to present and future generations."[2]

The global community has prior experience addressing the catastrophic risks of scientific and technological developments through international governance. In particular, AI researchers and policy experts have suggested that there might be lessons learned from the **history of nuclear stability** that could be applied to AI governance. Some developers of AI, such as Altman and others, have turned toward the **history of international nuclear nonproliferation agreements** for lessons in AI governance because their companies seek to balance the competing interests of market-based economics with those of collective public safety.[3]

Researchers, technology developers, and others have pointed to several analogies between AI and nuclear weapons that suggest the possibility of similar approaches to governance. Although these analogies are imperfect, there might be valuable lessons learned from reflecting on them, especially given how frequently those analogies have been publicly presented.[4]

For instance, one analogy focuses on the **dual-use nature of both technologies**, with each presenting the potential for national security and civilian applications, whether in nuclear energy production or AI-based drug research, among many other applications.[5] Another analogy draws a

[1] Sam Altman, Greg Brockman, and Ilya Sutskever, "Governance of Superintelligence," OpenAI blog, May 22, 2023.

[2] United Nations, "International Community Must Urgently Confront New Reality of Generative, Artificial Intelligence, Speakers Stress as Security Council Debates Risks, Rewards," July 18, 2023.

[3] Altman, Brockman, and Sutskever, 2023.

[4] For more on the value and disvalue of leveraging analogies to the history of governance for AI, see Michael J. D. Vermeer, *Historical Analogues That Can Inform AI Governance*, RAND Corporation, RR-A3408-1, 2024.

[5] Under the now revoked Executive Order 14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," the term *dual-use foundation model* is defined as "an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic

parallel between **the underlying materials needed for the technologies**, which remain highly specialized with restricted access (i.e., uranium or plutonium for nuclear energy production and computer chips to train and deploy advanced AI models).[6] Yet another analogy highlights the **rapid scientific advancements** of nuclear and AI technologies, which have produced an arms-race dynamic for countries to achieve advanced capabilities before their competitors.[7] Related is the **shared potential for catastrophic harm** that crosses borders, rendering regulation confined to national territories inadequate at best.[8] Such analogies are especially salient because tensions continue to escalate between global superpowers, which increases both the risks of nuclear instability and the competition to develop the most-advanced—and risky—AI systems.

Focus of This Paper

In this paper, we explore these issues by **examining key events in the history of nuclear nonproliferation agreements, having the goal to understand what lessons might be gleaned for AI governance**. We focus on two particular historical episodes in the development of the global nuclear governance regime: the Baruch Plan in 1946—an unsuccessful attempt to establish an international control regime surrounding nuclear weapons—and the more successful Treaty on the Non-Proliferation of Nuclear Weapons, commonly known as the Non-Proliferation Treaty (NPT), in 1968.[9] We seek to identify insights for establishing a similarly comprehensive and potentially even coercive governance regime for AI. Our review offers several insights for AI governance, although we ultimately find that there are significant hurdles that make proposed international approaches for AI governance that are directly modeled on this history unlikely to succeed. Nonetheless, this paper is intended as a provocative historical exploration to help lay the groundwork for future analysis and action.

By making these comparisons, we consider primarily AI governance proposals that suggest the creation of a global governance regime that is designed to oversee some or all aspects of AI.

security, national public health or safety, or any combination of those matters" (Executive Order 14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Executive Office of the President, October 30, 2023, p. 75194). Mauricio Baker, "Nuclear Arms Control Verification and Lessons for AI Treaties," arXiv, arXiv:2304.04123, April 8, 2023; Simon Chesterman, "Weapons of Mass Disruption: Artificial Intelligence and International Law," *Cambridge International Law Journal*, Vol. 10, No. 2, December 2021; Matthijs M. Maas, "How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons," *Contemporary Security Policy*, Vol. 40, No. 3, February 2019; Yonadav Shavit, "What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training Via Compute Monitoring," arXiv, arXiv:2303.11341, May 30, 2023; Kevin Klyman and Raphael Piliero, "AI and the A-Bomb: What the Analogy Captures and Misses," *Bulletin of the Atomic Scientists*, September 9, 2024.

[6] Baker, 2023; Chesterman, 2021; Dylan Matthews, "AI Is Supposedly the New Nuclear Weapons—but How Similar Are They, Really?" Vox, June 29, 2023; Shavit, 2023; Mike Watson, "IAEA for AI? That Model Has Already Failed," *Wall Street Journal*, June 1, 2023.

[7] Maas, 2019; Matthews, 2023; Waqar Zaidi and Allan Dafoe, "International Control of Powerful Technology: Lessons from the Baruch Plan for Nuclear Weapons," Centre for the Governance of AI, March 2021.

[8] Baker, 2023; Divyansh Kaushik and Matt Korda, "Panic About Overhyped AI Risk Could Lead to the Wrong Kind of Regulation," Vox, July 3, 2023; Klyman and Piliero, 2024; Maas, 2019.

[9] The Baruch Plan, presented to the United Nations Atomic Energy Commission, June 14, 1946; Treaty on the Non-Proliferation of Nuclear Weapons, signed at London, United Kingdom; Moscow, Russia; and Washington, D.C., on July 1, 1968.

Suggestions for a global AI governance regime are often incomplete and high-level and often made in the media by prominent figures in the AI industry but do not necessarily propose a comprehensive vision for what such a governance regime might entail.[10] Although we provide an illustrative overview of global AI governance proposals in the appendix, in this paper, we do not seek to comprehensively critique such statements for failing to specify an exact vision for global AI governance. Instead, we seek to illustrate the complications that come with comparisons to nuclear weapons and calls for global AI governance models that are based on the governance of nuclear weapons. Therefore, we hope to contribute to the discussion of how global AI governance should proceed by demonstrating how the analogy to nuclear weapons and such instruments as the NPT are, as of this writing, flawed in ways that make building such governance regimes as an NPT for AI difficult.

Our approach is to examine several key moments in the global governance of nuclear weapons, analyze the conditions that led to the failure or success of those attempts to global nuclear weapons, and illustrate a few key lessons those historical moments that might be helpful for attempts to govern AI. We then extract key lessons for the future governance of AI, whether by indicating pitfalls that those seeking to build a governance regime for AI should avoid or by analyzing the global conditions that made nuclear governance possible that might (or might not) exist for AI.

We do not attempt to provide a comprehensive comparison between all of nuclear history and AI; such a project would be of much greater scope than this paper. This publication is not intended to be a definitive analysis of nuclear history and its lessons for AI but rather to expand the understanding of the limitations of the nuclear analogy for AI. Therefore, our conclusions are necessarily limited by its selected case studies, and there might be additional historical evidence that further complicates the historical picture and provides additional lessons for AI governance.

The bulk of this analysis was done in the summer and fall 2024. Additional updates were done in spring 2025 to reflect advances in AI development.

The paper is organized as follows. In the next section, we review the motivation for and ultimate failure of the Baruch Plan, and in the third section, we describe the events leading to the successful NPT. In the penultimate section, we describe several challenges with proposals for AI governance modeled on this history, and we conclude with a summary of the lessons learned for future governance of AI.

The History of Nuclear Stability: The Failure of the Baruch Plan

We begin with a brief discussion of an unsuccessful attempt to establish a global nuclear governance regime.

In June 1946, U.S. diplomats made an astonishing offer to the UN: a proposal that the United States would relinquish its monopoly on nuclear weapons to an international body. Although most of the international community hailed this offer—known as the *Baruch Plan*—as magnanimous, Soviet officials perceived it as a bad-faith attempt to harm the interests of the Union of Soviet Socialist Republics (USSR) and rejected it. Historians acknowledge that Soviet diplomats had good reason to doubt that the United States would ever go through with the plan and that, even if the U.S. plan had

[10] For example, see Altman, Brockman, and Sutskever, 2023.

been made in earnest, the USSR could not have been reasonably expected to accept it. Moreover, it appears that the United States might have impaired the prospects for international cooperation to limit the proliferation of nuclear weapons by making the perfect the enemy of the good.

An Attempt to Rein in the Dangerous Potential of Nuclear Technology

The Baruch Plan emerged in the aftermath of World War II because the atomic bombings of Hiroshima and Nagasaki elicited a flurry of speculation among atomic scientists, government officials, and ordinary citizens that the new technology of nuclear weapons threatened to destroy civilization if nuclear weapons were not brought under control.[11] In November 1945, U.S. President Harry S. Truman, British Prime Minister Clement Attlee, and Canadian Prime Minister Mackenzie King issued a statement in which they declared that they wished to prevent the use of nuclear energy for destructive purposes. To this end, they called for a commission in the UN that would study how to accomplish this idealistic goal and determine how nuclear technology could be used for peaceful ends.[12] Although Soviet diplomats perceived this call as an attempt to pressure the Soviet Union, they elected to cooperate for tactical reasons, and the USSR cosponsored a resolution at the first session of the UN General Assembly in January 1946 that created the United Nations Atomic Energy Commission (UNAEC).[13]

Under intense time pressure, the U.S. government issued a report in March 1946, articulating a bold proposal to place atomic energy under international control.[14] **This report, which came to be known as the Acheson-Lilienthal Plan, divided the exploitation of nuclear energy into "dangerous" and "non-dangerous" activities and proposed that all of the former be placed under the control of an international agency.**[15] The plan delineated non-dangerous activities very narrowly, including such activities as the use of artificial radioisotopes for nuclear medicine but not including anything that might conceivably facilitate the production of weapon-grade fissile materials. Nation-states would not be permitted to mine their own uranium and thorium; instead, the envisioned international authority would do so when and how it saw fit.

Shortly before the publication of the Acheson-Lilienthal Plan, the revelation that the Soviet Union had spies in the Manhattan Project impelled the Truman administration to take a harsher line toward Moscow, in part because of the risk that the President's domestic political rivals would use atomic espionage as a cudgel to attack the Democratic Party.[16] To counter accusations that President Truman was being naïve about Soviet intentions, he named Bernard Baruch to lead the U.S.

[11] See Chapter 3 in Paul Boyer, *By the Bomb's Early Light: American Thought and Culture at the Dawn of the Atomic Age*, Pantheon, 1985.

[12] Michael D. Gordin, *Red Cloud at Dawn: Truman, Stalin, and the End of the Atomic Monopoly*, Farrar, Straus and Giroux, 2009, p. 39.

[13] David Holloway, *Stalin and the Bomb: The Soviet Union and Atomic Energy, 1939–1956*, Yale University Press, 1994, p. 161.

[14] Campbell Craig and Sergey Radchenko, *The Atomic Bomb and the Origins of the Cold War*, Yale University Press, 2008, pp. 120–121.

[15] The report defined three activities as dangerous: the supply of raw materials (uranium and thorium), the conversion of these materials into weapon-grade fissile materials, and the design and fabrication of nuclear weapons.

[16] Craig and Radchenko, 2008, pp. 121–122.

delegation to UNAEC. One of Baruch's conditions for accepting the position was that he be permitted to put his own stamp on the Acheson-Lilienthal Plan. Truman's memoirs suggest that he selected Baruch to lead negotiations because Baruch held significant sway in the Senate that might help overcome congressional opposition to the plan, although there was significant concern about his appointment from others in Washington who considered him too ornery and unfamiliar with nuclear technology to lead the negotiations.[17]

The plan that Baruch presented at UNAEC on June 14, 1946, followed the broad outlines of the Acheson-Lilienthal Plan.[18] Under the Baruch Plan, the United States would eliminate its nuclear arsenal and provide nuclear technology to other countries. However, other nations would be required to forgo the means to develop nuclear weapons and agree to a system of inspections to ensure compliance. An international body would also be established to support nuclear technology, and this body would have a monopoly on mining uranium and thorium, the refining of those elements, and the construction and operation of nuclear power plants. However, **the Baruch Plan stipulated that the international control regime would have to be established and functioning before the United States would surrender its nuclear weapons**. While such a regime was being established and before the United States would surrender its weapons, all countries would be obligated to open their borders for inspection to determine their nuclear resources and ongoing nuclear activities.[19] In short, the USSR (and other nations) would have to pledge to cease any nuclear weapon development and open their borders to inspections, whereas the United States would, at least temporarily, maintain its nuclear arsenal.

On June 19, 1946, Soviet Minister of Foreign Affairs Anatolii Gromyko responded to the Baruch Plan with a Soviet counterproposal, which proposed that the first step ought to be an international convention banning the production, stockpiling, and use of nuclear weapons. All existing atomic bombs would be destroyed within three months of this convention entering into force. Signatory states would enact national legislation to establish punishments for those that violated the convention.[20] Unlike the Baruch Plan, which called for an unprecedented degree of international control, the Gromyko Plan did not include any form of international control or oversight.

Baruch was eager to bring the U.S. proposal up for a vote before the end of 1946, when the composition of the UN Security Council would give greater representation to the Soviet Union's Eastern European client states.[21] Voting on the proposal was delayed because the Soviets made counterproposals for international inspections during negotiations that drew interest from the United Kingdom and France. However, the earnestness of such proposals was unclear, and because of U.S. pressure, all members of the UN Security Council, with the exception of the USSR and Poland, voted

[17] Gordin, 2009, pp. 63–65.

[18] Albeit with a few key additions. The Acheson-Lilienthal Plan was silent about the sanctions that would be imposed by the international community on malefactors that engaged in prohibited nuclear activities. The Baruch Plan stipulated that violators would be punished and, furthermore, that a UN Security Council veto would not be applicable in this instance.

[19] Gordin, 2009, p. 52; Gregg Herken, *The Winning Weapon: The Atomic Bomb in the Cold War, 1945–1950*, Alfred A. Knopf, 1980, pp. 163–164.

[20] Holloway, 1994, pp. 161–162.

[21] Herken, 1980, p. 172.

to endorse the Baruch Plan.[22] Negotiations in the UNAEC continued until July 1949 amid deepening U.S.-Soviet tensions. The following month, the USSR tested its first nuclear weapon, ending the U.S. atomic monopoly.[23]

The Reasons the Baruch Plan Failed

History suggests several implications for efforts to build cooperation on international governance that are relevant for AI. In particular, the case of the Baruch Plan suggests that **schemes for international control of dual-use technologies are only likely to succeed if all relevant stakeholders perceive them to be earnest and equitable** and that proposals that perpetuate one nation's advantage might be perceived as cynical gambits and ultimately can make long-term cooperation significantly more difficult than it could have been otherwise.

Although historical views of the Baruch Plan differ, **it is clear that the Baruch Plan was not perceived as earnest and equitable by all relevant stakeholders**. Although some U.S. accounts portray the plan as an earnest offer by the United States to give up the crown jewel of its defense technology for the greater good, which was rejected unreasonably by a truculent Soviet Union,[24] other accounts, including essentially all Soviet and Russian ones,[25] characterized the U.S. proposal as an exercise in *Machtpolitik* that was intended to further U.S. interests.[26] The Soviet position contended that the Baruch Plan was formulated and advocated with the primary goal of reinforcing U.S. national security interests, whether by compelling the Kremlin to embarrass itself by rejecting an ostensibly generous offer or by imposing terms that granted advantages to Washington if the Soviet Union accepted them.[27] A third school of thought argued that the Baruch Plan resulted from primarily domestic U.S. political calculations rather than the imperatives of international competition. In the aftermath of the atomic espionage revelations in early 1946, the Baruch Plan arguably constituted a political compromise offering the appearance of a balance between the contradictory objectives of preserving U.S. advantage and embracing international control.[28]

Even if Baruch and his colleagues made their proposal for the international control of nuclear energy in good faith, **there are compelling reasons to suspect that the United States would not have followed the terms articulated by Baruch at the UN in June 1946**. Giving up the U.S. nuclear arsenal would not have been acceptable to U.S. domestic constituencies. Opinion polls showed that a

[22] Gordin, 2009, p. 158.

[23] Holloway, 1994, pp. 213–218.

[24] Craig and Radchenko, 2008, p. 125.

[25] Gordin, 2009, p. 53; V. A. Tarasenko, *The Atomic Problem in the External Relations of the U.S.A.* [Атомная проблема во внешней политике США] Shevchenko Kyiv State University Press [Izd-vo Kievskogo gosudarstvennogo universiteta im. T. G. Shevchenka], 1958.

[26] Patrick Maynard Stuart Blackett, *Fear, War, and the Bomb: Military and Political Consequences of Atomic Energy*, Whittlesey House, 1949, pp. 143–194; Craig and Radchenko, 2008, pp. 125–127.

[27] Blackett, p. 192.

[28] Boyer, 1985. p. 54; Herken, 1980, pp. 173–174. The arguments of the *Machtpolitik* interpretation and the domestic political interpretation are not mutually exclusive because they differ according to the perceived relative influence of domestic concerns versus international ones that drove the formulation of the U.S. policy that led to the Baruch Plan. Obviously, both domestic and international concerns played a role, and each had a varying degree of influence on individuals and across time.

sizable majority (72 percent) of the U.S. population was loath to give up the "atom bomb secret" as of August 1946, even though 69 percent expressed support for international control of atomic energy.[29] More important, U.S. officials increasingly perceived nuclear weapons as a counterweight to the USSR's large conventional forces.[30]

Even if we presume that the Baruch Plan was put forward in good faith, **Baruch and his colleagues conducted their diplomacy in a way that could be expected to antagonize their Soviet counterparts.** Although Baruch insisted in his public speeches that his plan was an opening bid for further negotiations, he demonstrated no actual willingness to make meaningful concessions to the Soviet position.[31] Soviet participants in the negotiations assumed that the United States would never willingly surrender its nuclear weapons.[32]

The Baruch Plan was engineered to favor U.S. interests rather than Soviet ones, a fact that Western observers who were sympathetic to Moscow immediately pointed out and that others recognized more belatedly.[33] Moreover, the USSR was determined to develop nuclear weapons and expected to have them in a few years. The ostensible U.S. lead in technology and resources that Baruch assumed would provide the United States with immense negotiating leverage was far smaller than he imagined.[34] In hindsight, the Baruch Plan left a problematic legacy for later attempts to establish an international cooperation to limit nuclear weapon proliferation and reduce the risk of nuclear war. Although the Baruch Plan attained a significant, if temporary, diplomatic and propaganda victory, it antagonized the USSR.

It might be the case that, over the longer term, the U.S. government was worse off for having staked its legitimacy on a maximalist international control solution instead of exploring more-attainable partial solutions. Significantly, the international arrangements that later emerged to control nuclear materials under International Atomic Energy Agency (IAEA) and the NPT auspices ultimately resembled the proposals that Soviet negotiators put forward in 1947, which U.S. officials completely rejected at the time, which suggests that these structures were the only ones to which all parties could, in the end, agree as being the basis for nonproliferation.[35] The legacy of the Baruch Plan might have made it more difficult to reach agreement on these matters and delayed the date when such agreements would be enacted.

In short, the Baruch Plan might have been worse than nothing. Therefore, the Baruch Plan primarily serves as an example of what **not** to do when designing and negotiating an international control regime for a dangerous new technology. This history lesson suggests that, when it comes to AI, attempts to negotiate global governance from a position of strength or technological advantage

[29] Herken, 1980, p. 180.

[30] Herken, 1980, pp. 178–179.

[31] Craig and Radchenko, 2008, p. 126; Herken, 1980, Ch. 9.

[32] Holloway, 1994, pp. 161–165.

[33] Blackett, pp. 184–186; Henry A. Wallace, "From the Letter to the President," *Bulletin of the Atomic Scientists*, Vol. 2, Nos. 7–8, 1946, pp. 2–3.

[34] Gordin, 2009, Ch. 4; Holloway, 1994, pp. 220–223. In 1946, General Leslie Groves, who directed the Manhattan Project, assured Baruch that the USSR was 20 years away from having an atomic bomb of its own; in actuality, the first Soviet nuclear test took place only three years later in 1949. See Herken, 1980, p. 168.

[35] Bertrand Goldschmidt, *A Forerunner of the NPT? The Soviet Proposals of 1947*, International Atomic Energy Agency, Vol. 28-1, March 1986, pp. 58–64.

would be difficult, particularly if other states perceive such negotiating offers as being intended to lock in a U.S. advantage.

The History of Nuclear Stability: The Non-Proliferation Treaty and Strategic Stability

Roughly 20 years after the proposal of the Baruch Plan, the NPT was negotiated, signed in 1968, and entered into force in 1970. The NPT, which has been of particular interest for AI policymakers as a potential model for international agreements to limit proliferation of powerful AI systems, came about through a confluence of geopolitical factors and specific incentives that brought both the United States and the USSR to the bargaining table in 1960s.[36]

The Geopolitical Climate in the Early Cold War

The geopolitical climate created by the Cold War delayed the pursuit of a nonproliferation agreement. This context was characterized by several factors.

U.S. exclusive control over nuclear weapons ended shortly after the failure of the Baruch Plan when the Soviets successfully tested a nuclear weapon in 1949.[37] The USSR was able to overcome perceived technical obstacles to fielding its nuclear weapons faster than the United States anticipated, putting to bed the notion of maintaining U.S. nuclear exclusivity.

The early Cold War was characterized by intense debate over whether and how nuclear weapons should be deployed, whether a nuclear conflict was winnable, and if so, how to win such a conflict.[38] Early nuclear delivery systems, particularly bombers, were also potentially vulnerable to attack and could be intercepted or destroyed on the ground by a successful surprise attack.[39]

The competitive atmosphere of the early Cold War was not conducive to the development of a nonproliferation agreement between the United States and the USSR. **Official nuclear policy focused primarily on superpower confrontation, and both sides sought to develop sufficient nuclear arsenals to deter the other.** Eisenhower's Atoms for Peace program, which promised nuclear assistance for peaceful purposes and the creation of the IAEA in 1957, did suggest a potential architecture for nonproliferation but was not immediately acted on.[40]

[36] Andrew J. Coe and Jane Vaynman, "Collusion and the Nuclear Nonproliferation Regime," *Journal of Politics*, Vol. 77, No. 4, January 2015.

[37] William Burr, ed., "Detection of the First Soviet Nuclear Test, September 1949," National Security Archive, September 9, 2016.

[38] Herman Kahn, *On Thermonuclear War*, Princeton University Press, 1960, is the most prominent example of playing out how a nuclear war might be won.

[39] Albert Wohlstetter, Fred Hoffman, R. J. Lutz, and Henry S. Rowen, *Selection and Use of Strategic Air Bases*, RAND Corporation, R-266, 1954, discusses the vulnerabilities of nuclear bombers and how to address these vulnerabilities to increase the force's resilience.

[40] Dwight D. Eisenhower, "Atoms for Peace Speech," address to the 470th Plenary Meeting of the United Nations General Assembly, December 8, 1953.

Evolving Perceptions of Risk

Several technical and political developments were necessary to increase the salience of nuclear proliferation to third-party states and create the conditions for a U.S.-USSR agreement. One key factor was a growing awareness of the risks inherent to the nuclear status quo.

The issue of **nuclear proliferation to additional countries rose in prominence** as the United States and USSR developed their nuclear forces. At first, development of nuclear weapons by third parties was given comparatively less attention than attending to the nuclear balance between the two great powers.[41] Over time, more attention began to be paid to third-party attempts at developing nuclear weapons when additional nations nuclearized in the 1950s and 1960s.[42] For example, a study in the 1960s discusses the "Nth country problem," in which third parties that are outside a major alliance and have nuclear weapons might catalyze nuclear war, and Lyndon Johnson and others worried after the nuclearization of China that "proliferation cascades" could break out if too many nations were allowed to nuclearize.[43] These specific concerns were accompanied by broader strategic thinking that identified nonproliferation as a strategy that could both enhance U.S. leverage over nonnuclear nations and reduce the chance of a nuclear war that targeted the United States.[44]

Concern about the nuclear threat was heightened by the rapid advance of nuclear technology, including the development of increasingly powerful weapons, such as the hydrogen bomb.[45] Early on, U.S. policymakers believed it would be decades before the Soviet Union produced a nuclear bomb, an assumption of which they were quickly disabused when the USSR acquired nuclear weapons in 1949.[46] However, policymakers continued to think that nuclear weapon production would require a level of significant technical and industrial complexity that most nations did not possess.[47] This thought process changed when uranium enrichment using gas centrifuge technology, which significantly reduced the technical barriers for middle powers to produce their nuclear arsenals, was developed.[48] These technologies increased the potential for nuclear proliferation, an issue that concerned U.S. and British policymakers in the lead-up to the NPT's negotiation and ratification.[49] Furthermore, nuclear weapons increased in power vastly over the same time, particularly the development of the hydrogen bomb, which augmented the danger posed by nuclear weapons.[50]

[41] Roland Popp, Liviu Horovitz, and Andreas Wagner, eds., *Negotiating the Nuclear Non-Proliferation Treaty: Origins of the Nuclear Order*, Routledge Taylor & Francis Group, 2017, p. 10.

[42] Francis J. Gavin, "Strategies of Inhibition: U.S. Grand Strategy, the Nuclear Revolution, and Non-Proliferation," *International Security*, Vol. 40, No. 1, Summer 2015, p. 21.

[43] Gavin, 2015, p. 21.

[44] Gavin, 2015, pp. 21–24.

[45] Popp, Horovitz, and Wagner, 2017, p. 11.

[46] R. Scott Kemp, "The End of Manhattan: How the Gas Centrifuge Changed the Quest for Nuclear Weapons," *Technology and Culture*, Vol. 53, No. 2, April 2012, p. 273.

[47] Kemp, 2012, p. 273.

[48] Popp, Horovitz, and Wagner, 2017, p. 14.

[49] John Krige, "The Proliferation Risks of Gas Centrifuge Enrichment at the Dawn of the NPT: Shedding Light on the Negotiating History," *Nonproliferation Review*, Vol. 19, No. 2, July 2012, p. 220.

[50] Center for Arms Control and Non-Proliferation, "Fact Sheet: Thermonuclear Weapons," November 18, 2022.

In addition, near misses in nuclear war threw the high stakes of nuclear weapon brinksmanship into sharp relief and induced both sides to consider alternatives to maximalist confrontation. The Cuban Missile Crisis, in particular, demonstrated the risks of nuclear confrontation through the deployment of such weapons in the territories of third powers. The Cuban Missile Crisis and other close calls did not lead directly to nonproliferation negotiations but clarified the potential costs of nuclear brinksmanship, helping move the United States and USSR away from maximalist confrontation and toward policies that could reduce the chance of nuclear confrontation.[51] This would in turn create space for discussion of and negotiation toward nonproliferation.

Furthermore, developments of missile technology reinforced mutually assured destruction (MAD) by reducing response time and increasing the difficulty of eliminating an opponent's nuclear arsenal.[52] **The end result, which officials in the United States and the USSR recognized, was the development of a nuclear stalemate based on MAD and the resulting theories of nuclear stability.**[53]

Nuclear Weapons as Potential Tools of International Stability and Conflict Prevention

The interconnected theories of deterrence and MAD were key to the development of a stable global nuclear order. Although these concepts are now deeply embedded in the understanding of nuclear weapons, during the early Cold War, nuclear weapons were a new technology that did not have developed norms.[54] Early Cold War thinking on nuclear war included a lively debate on how a nuclear war might be won, who should have authority over such weapons, and when those weapons should be deployed.[55]

These debates were resolved over time, particularly when technical innovations increased the survivability of each superpower's nuclear deterrent and second-strike capability in a nuclear war.[56] Both sides developed mature second-strike capabilities, and **theorists developed the concepts of MAD and nuclear stability, which understood nuclear weapons as potential tools of international stability and conflict prevention because each superpower's nuclear arsenal guaranteed that neither side could gain a sufficient first-mover advantage in a nuclear conflict**. These concepts were accepted by the United States and the USSR, and both states' interests in winning a nuclear war gave way to a desire to prevent the emergence of new nuclear powers and avoid destructive nuclear war. In

[51] Mary Olney Fulham, "Ask the Experts: The 60th Anniversary of the Cuban Missile Crisis," Nuclear Threat Initiative, October 13, 2022.

[52] Popp, Horovitz, and Wagner, 2017, p. 11.

[53] Popp, Horovitz, and Wagner, 2017, p. 11.

[54] Charles H. Fairbanks, Jr., "MAD and U.S. Strategy," in Henry D. Sokolski, ed., *Getting MAD: Nuclear Mutual Assured Destruction, Its Origin and Practice*, U.S. Army War College Press, 2004, discusses the development of theories of MAD and other theories of nuclear deterrence since the 1950s.

[55] Fairbanks, 2004, p. 137.

[56] William Burr and David Rosenberg, "Nuclear Competition in the Age of Stalemate, 1963–1975," in Melvin P. Leffler and Odd Anne Westad, eds., *The Cambridge History of the Cold War*, Vol. II, *Crises and Détente*, Cambridge University Press, 2010, discusses the development of nuclear stalemate and the technical innovations leading to this state of affairs.

this context, discussions about arms control, including nonproliferation, became more plausible between both superpowers.

The development of a strategy of nuclear stalemate also incentivized both countries to begin to consider nuclear nonproliferation more closely to preserve stability between both sides. While the consensus surrounding nuclear stalemate developed, **there was increasing concern about proliferation because of "nuclear plenty," when untrustworthy nations gained access to nuclear weapons and were potentially capable of destabilizing the global balance of power.**[57] The technical advances discussed previously also increased the perceived threat from nuclear proliferation because more nations were judged capable of developing a nuclear arsenal.

The Benefits of Collusion Between the Two Superpowers

When the international environment became more conducive to nonproliferation negotiations, **the United States and USSR began to see potential benefits from colluding to limit the spread of nuclear weapons.**[58] The continued proliferation of nuclear weapons, particularly to nations outside each superpowers' close partners, increased concerns that nuclear weapons would be developed by untrustworthy actors outside the Cold War alliance system. This proliferation, in the context of strategic stability, was a potentially destabilizing force that could upset the nuclear status quo between the United States and the USSR, which each possessed a full-fledged and mature nuclear force. These concerns were particularly acute in the USSR, which feared that West Germany might develop or be transferred nuclear weapons by its North Atlantic Treaty Organization (NATO) allies.[59]

The Role of Extended Deterrence

In this context, the concept of extended deterrence became important, whereby the United States discouraged attacks on its partners by guaranteeing that attacks on their territory will be met by a nuclear response.[60] The credibility of extended deterrence was a significant issue throughout the Cold War. Questions arose concerning the credibility of commitments by the United States to deploy its nuclear weaponry in response to an attack on an ally, which might, in turn, provoke an attack against the United States.[61] The United States attempted to answer concerns about credibility in several ways; for instance, in the early Cold War, the United States did not oppose the development of independent, European nuclear arsenals in the United Kingdom and France.

The United States also sought to address concerns about extended deterrence through multilateral mechanisms. For example, the United States proposed a NATO Multilateral Force

[57] Popp, Horovitz, and Wagner, 2017, p. 10.

[58] See, for example, Coe and Vaynman, 2015.

[59] Johathan Hunt, "'If One Tightens the Screw to the Limit . . . One Might Strip the Thread': Soviet Defenses of the Nuclear Non-Proliferation Treaty," Sources and Methods blog, September 7, 2023; Andreas Lutsch, "The Federal Republic of Germany and the NPT, 1967–1969," Sources and Methods blog, January 29, 2024.

[60] Michael J. Mazarr, *Understanding Deterrence*, RAND Corporation, PE-295-RC, April 2018, p. 3.

[61] Mazarr, 2018.

(MLF), a plan to create a fleet of NATO-commanded nuclear-armed ships and submarines.[62] The MLF was originally proposed to serve as a shared command-and-control mechanism to address concerns about the credibility of U.S. commitments to its European allies.[63] However, the USSR strenuously opposed the MLF because of concerns about it granting nuclear weapons to West Germany, considering the aggression by the Germans against the USSR during World War II. The USSR argued that the creation and arming of such a force should be prohibited under any nonproliferation treaty.[64] Indeed, concerns about West German nuclear access, in particular, made nonproliferation a concrete issue for the USSR, incentivizing the Soviets to bargain seriously regarding the NPT as a way to permanently prevent West German nuclear proliferation.[65]

However, there was the potential for compromise between the United States and the USSR.[66] By 1966, the Soviets received assurances that the United States would not transfer nuclear weapons to West Germany, which satisfied the USSR that the Germans would not receive nuclear weapons.[67] This no-transfer provision would become a central component of the NPT. However, the United States would, under the treaty's terms, be allowed to station nuclear weapons in West Germany and other allied nations, provided that those weapons were under the sole control of U.S. personnel, thereby ensuring the United States could continue to deter potential Soviet aggression.[68] This compromise satisfied the Soviet fear of West Germany developing its own nuclear armaments and allowed the United States to advance its broader interest in nuclear nonproliferation.

Ultimately, extended deterrence served to support nonproliferation by removing the need for allies and partners of the United States to develop their own nuclear weapons. The end result was to reduce the proliferation of nuclear weapons among U.S. partners, which thought that they could rely on the nuclear umbrella that the United States provided.[69]

[62] William Burr, ed., "Preoccupations with West Germany's Nuclear Weapons Potential Shaped Kennedy-Era Diplomacy," National Security Archive, February 2, 2018a; James B. Solomon, *The Multilateral Force: America's Nuclear Solution for NATO (1960–1965)*, Naval Academy, May 4, 1999.

[63] Solomon, 1999.

[64] Burr, 2018a.

[65] Concerns about Chinese nuclear capabilities also played a role in incentivizing the United States and the USSR to negotiate the NPT more seriously. See Hunt, 2023. It should be noted that the United States had essentially promised the Soviets that West Germany would not have nuclear weapons in 1963, and therefore, West Germany's potential nuclearization might have been less pressing for the Soviets because this issue might have been considered to be mostly resolved.

[66] George Bunn, "The Nuclear Nonproliferation Treaty: History and Current Problems," Arms Control Association, December 2003.

[67] William Burr, ed., "The Nuclear Nonproliferation Treaty and the German Nuclear Question Part II, 1965–1969," National Security Archive, March 21, 2018b.

[68] Bunn, 2003.

[69] Justin Anderson, Jeffrey Larsen, and Polly Holdorf, *Extended Deterrence and Allied Assurance: Key Concepts and Current Challenges for U.S. Policy*, U.S. Air Force Institute for National Security Studies, September 2013.

Why Both Sides Were Willing to Compromise to Establish a Nuclear Governance Regime

We now return to the larger question underlying this discussion, which is why the United States and USSR were willing to compromise to reduce the threat of nuclear nonproliferation. Several factors help explain the successful negotiation when previous attempts at nonproliferation had been less successful.

Several theories have been put forward to explain why the United States and the USSR agreed to the NPT. However, before we discuss those theories, it should be noted that the reasons are not exclusive of each other as being motives for establishing the NPT but, in fact, were potentially synergistic in motivating the parties to come to the table.

One theory considers the superpowers and their European allies to share an interest in codifying the nuclear order in Europe and accepting the Cold War settlement of territory in Europe. This approach represented a change from earlier in the Cold War, when both sides did not perceive the European status quo as mostly settled and, therefore, thought that they could successfully out-compete their opponents on the continent.

A second theory emphasizes that, although German nuclearization was a major issue, both superpowers also had significant concerns about proliferation worldwide, which motivated them to come to the table to create a "grand bargain" to contain nuclear proliferation.[70] However, both theories support the notion that the United States and the Soviet Union had grounds for collusion that would preserve their nuclear strength while limiting the ability of other third powers to join the nuclear club and potentially destabilize the global balance of power.[71]

Another overarching reason supporting the willingness to compromise is that both sides ultimately accepted the value of strategic stability. The definition of *strategic stability* is notoriously difficult to pin down, and many authors have offered their formal definitions of the term.[72] This said, one definition that captures the key components of nuclear strategic stability is that **no side feels that using nuclear weapons could result in a better outcome than continued nuclear deterrence and pursuit of national security objectives by nonnuclear means**. The development of strategic stability between the United States and the USSR created incentives to take nuclear nonproliferation seriously in order to preserve the balance of power.[73]

Strategic stability supports nonproliferation because, **under conditions of strategic stability, nuclear parties are incentivized to prevent proliferation that might upset the strategic stability equilibrium between each party and its competitors**. The stable equilibrium of strategic stability also relied on the development of resilient, second-strike capabilities by the superpowers and the acceptance of MAD and deterrence as the frameworks for the usage of nuclear weapons. In the early Cold War period, nuclear overmatch still seemed possible, and therefore, neither side could trust in a stable equilibrium because the other side might seek to gain a decisive nuclear advantage. When it

[70] Leonard Weiss, "Nuclear-Weapon States and the Grand Bargain," Arms Control Association, December 2003.

[71] See, for example, Coe and Vaynman, 2015.

[72] Elbridge A. Colby and Michael S. Gerson, eds., *Strategic Stability: Contending Interpretations*, Strategic Studies Institute and U.S. Army War College Press, February 2013.

[73] Popp, Horovitz, and Wagner, 2017, p. 10.

became clear that no such decisive advantage could be gained, strategic stability solidified, and both sides accepted that deploying nuclear weapons would not grant a sufficient advantage to justify their use.[74]

Offering Carrots and Sticks to Encourage Compliance

In addition, the NPT not only satisfied the interests of the nuclear superpowers at the time but also offered a mutually agreed-on means to encourage compliance that was acceptable to both sides. The NPT is sustained by offering both carrots, in the form of access to civilian nuclear technology for complying nations, and sticks, in the form of sanctions and other even military actions against states that attempt to develop nuclear weapons after the NPT came into force. The carrots provide the material incentives that keep nonnuclear nations (mostly) in the NPT, though such states as North Korea and Iran have still sought nuclear weapons to secure their own strategic interests, despite the requirements of the NPT and the significant costs imposed on both states as a result of their nuclear programs.

The success of these carrots and sticks has been reliant on coordination and cooperation of the nuclear-armed powers, which historically have continued to cooperate under the NPT to minimize the emergence of new nuclear-armed states. This success is because the incentives that drove these countries to negotiate the NPT in the first place have remained relatively static; each nuclear-armed power derives significant benefits from keeping the nuclear club small and maintaining nuclear stability between each other. **The stable benefit of the NPT to nuclear powers has undergirded the treaty's relative success and survival since its negotiation.**

Identifying Challenges in Establishing an International Governance Regime for AI from Past Experiences with Nuclear Weapons

The experience of creating the NPT, nuclear stability, and a nonproliferation regime that eventually evolved relies on policymakers having a clear understanding of (1) the underlying risks (e.g., nuclear weapons with catastrophic impact, underlying nuclear material), (2) what parties actually govern to reduce risk (i.e., states with nuclear weapons), and (3) what mechanisms will be used for governance (through nonproliferation and international mechanisms). **Would a similar sort of governance regime be possible to rein in the potential risks from powerful AI?**

In short, the answer is not yet. At this point, **the development of an AI governance approach modeled on the nuclear stability regime is hindered across all three areas described previously: disagreement about AI risks, the complex ecosystem of actors, and uncertainties about governance approaches.** The history of the NPT also suggests that global AI governance, particularly the sort that could prevent proliferation of the most powerful models, might occur only when major AI powers reach parity with each other and there are conditions of mutual vulnerability.

[74] Colby and Gerson, 2013, p. 202.

Because implementing international governance of AI could involve difficult political decisions that might be viewed by some stakeholders and the domestic public as negatively affecting national security and economic growth—and perhaps also set back innovation that could improve people's lives—these disagreements create significant practical barriers for developing an overarching catastrophic AI risk regime in the near and medium terms.

In the following sections, we describe several major challenges to governing the catastrophic risks of AI in a way that is modeled on the nuclear stability regime.

Disagreement About AI Risks

First, **there is persistent disagreement and uncertainty about the extent to which existing or even future AI technologies present extreme or catastrophic risks comparable to the destructive power of nuclear weapons.**

A growing cross section of notable technologists, policymakers, and researchers have noted that powerful AI poses catastrophic risks.[75] Researchers who suggest AI poses such risks have tended to focus on two types of risk: (1) misuse of AI, in which AI enables a human to develop and deploy a catastrophic weapon (e.g., a bio- or cyberweapon), and (2) rogue AI, in which a highly capable AI pursues objectives that are harmful to humanity, perhaps through deception or surreptitious self-replication.

Those who emphasize AI misuse risks included the Biden administration, whose now-revoked Executive Order 14110 on AI notes that AI might pose risks by "substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons."[76] Similarly, technologists and researchers have argued that AI might enable the design and use of biological weapons.[77]

However, AI companies seem to assess that the AIs they have developed so far do not provide the necessary uplift for catastrophic misuse. For instance, OpenAI's most recent published evaluation of GPT-4 finds that it provides "at most a mild uplift in biological threat creation accuracy" with an effect size that is not statistically significant.[78] Moreover, this evaluation considered only how AI provides the needed knowledge and did not consider other significant barriers to developing biological weapons, including accessing and synthesizing the necessary materials. Other research also finds that the deployed AI does not alleviate the bottlenecks that would limit biological or chemical weapon deployment.[79]

[75] See, for example, Center for AI Safety, "Statement on AI Risk," undated.

[76] Executive Order 14110, 2023, p. 75194.

[77] Dario Amodei, "Oversight of A.I.: Principles for Regulation," testimony before the U.S. Senate Judiciary Committee Subcommittee on Privacy, Technology, and the Law, July 25, 2023; Jason Matheny, "Challenges to U.S. National Security and Competitiveness Posed by AI," testimony presented before the U.S. Senate Committee on Homeland Security and Governmental Affairs, RAND Corporation, CT-A2654-1, March 8, 2023.

[78] Tejal Patwardhan, Kevin Liu, Todor Markov, Neil Chowdhury, Dillon Leet, Natalie Cone, Caitlin Maltbie, Joost Huizinga, Carroll Wainwright, Shawn (Froggi) Jackson, et al., "Building an Early Warning System for LLM-Aided Biological Threat Creation," OpenAI, January 31, 2024.

[79] Christopher A. Mouton, Caleb Lucas, and Ella Guest, *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study*, RAND Corporation, RR-A2977-2, 2024.

The risks of a rogue AI are similarly contentious, and AI researchers and policymakers debate the plausibility that the speculative scenarios familiar from science fiction will manifest. Some researchers who are closest to the technology express confidence that AI will never become uncontrollable, pointing to a likely plateau in AI performance (because of ever-increasing data needs, cost and energy demands of training runs, and limits to the existing architecture).[80] Of course, this is a lively debate, and others might, in turn, suggest that an AI superintelligence that is prone to deceive us is already here.[81] In general, evaluating whether an advanced AI system will pose catastrophic risks could be difficult. Standards for evaluation are still in development (even as the capabilities continue to grow), and independent researchers do not always have access to closed (i.e., protected proprietary) models.

Beyond the two major pathways to catastrophe of misuse and rogue AI, **some have discussed other ways that AI might result in severe harm to humanity**, including those that suggest that AI might make humans obsolete or that AI might bolster a totalitarian state that restricts fundamental human choices.[82] **However, these alternative pathways to catastrophe are the result of accumulating harms that compound rather than a sudden all-at-once event.** In this way, these pathways might be substantially different and subject to more complexity than the sort of nuclear conflict that motivated the nuclear stability regime. Furthermore, divergent pathways or risks might necessitate different forms of governance, and there might not be such a single treaty as the NPT that can account for this variety.

Others disagree that AI even presents catastrophic risks. These analysts reply that the discussion of catastrophic AI risks is an attempt to hype the technology and hide the real everyday risks, such as civil liberties violations, the concentration of corporate and political power, and disparate harm to vulnerable communities.[83] These perspectives contend that approaches to govern misuse and rogue AI risks are an attempt by the major technology players to distract regulators from ongoing risks, extend AI companies' dominance over the technology, and stifle competition or global development.

The upshot of this discussion is that **there is plurality of views about whether AI in fact poses catastrophic risk and about the possible pathways to an AI-based catastrophe**. Each of these different pathways might have different implications for what type of governance is necessary. If, indeed, a regime is developed to address a specific type of risky pathway, this might, as suggested by Liu, Lauta, and Maaset, "only afford future policymakers with a false sense of security" and thereby

[80] See, for example, Tammy Xu, "We Could Run Out of Data to Train AI Language Programs," *MIT Technology Review*, November 24, 2022.

[81] See, for example, Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," arXiv, arXiv:2303.12712, April 13, 2023; and Nitasha Tiku, "The Google Engineer Who Thinks the Company's AI Has Come to Life," *Washington Post*, June 11, 2022.

[82] Kelsey Piper, "Is It Time for a Pause?" Planned Obsolescence blog, March 30, 2023; Toby Ord, The Precipice: *Existential Risk and the Future of Humanity*, Bloomsbury Publishing, 2020.

[83] See, for example, Joy Buolamwini, *Unmasking AI: My Mission to Protect What Is Human in a World of Machines*, Random House, 2023; "Stop Talking About Tomorrow's AI Doomsday When AI Poses Risks Today," *Nature*, Vol. 618, No. 7967, June 2023; and Matteo Wong, "AI Doomerism Is a Decoy," *The Atlantic*, June 2, 2023.

discourage them from taking action toward addressing other risks.[84] Thus, these disagreements about the nature of AI risk might lead policymakers to address an incomplete set of risks.

The absence of a consensus around AI risks might more likely flatly discourage policymakers and AI firms from making any decisions that will be economically painful and politically challenging. It is important to remember that **if there is a plurality of divergent and outspoken positions on the nature of AI risk, those seeking to establish AI governance mechanism will be met with resistance from various constituents**.

In the context of nuclear risk, there were several concrete moments that helped to demonstrate the risks that nuclear weapons posed and galvanize a consensus for international action. For instance, the 1945 Trinity test provided a demonstration of the destructive nature of nuclear explosions, followed by many other tests whose successes were widely acknowledged (and feared) globally. The nuclear attacks on Hiroshima and Nagasaki similarly provided a horrific demonstration of these weapons. International action to develop global institutions was significantly spurred by the Cuban Missile Crisis, which helped catalyze global leaders to make the hard political decisions to commit to monitoring regimes and any provision of civilian use. The global institutions developed to manage nuclear risks, such as the IAEA, are focused on a manageable set of well understood risks. In comparison, consensus on the risks from AI is still lacking.

In addition, in contrast to nuclear weapons, **there is no AI stability or AI stalemate yet**, either between countries that host firms developing advanced AI or among the firms. Indeed, the underlying technology is still perhaps too immature and its potential too uncertain for a clear understanding on how a powerful AI might be used to advance national security or threaten the interests of other countries. There is also no strategic balance in AI, in which the use of AI is discouraged by potentially large negative consequences. Instead, many policymakers see significant benefits from increasing the deployment of AI across more fields for both strategic and economic advantages.

AI capabilities are developing quickly, and in the near term, there might be a demonstration of a capability that will help rally attention and create conditions for stability (or instability).[85] Perhaps, key stakeholders will eventually foster consensus on how AI capabilities can result in severe harm. However, until then, **these disputes will derail difficult action, hinder the consensus on risks, and hamper the mutual recognition of the value of controlling AI** that nuclear history suggests was necessary to reach global arrangements for nuclear stability.

Complex Ecosystem of Players

Another challenge in establishing an AI governance regime involves the number of actors that would need to agree on the basic arrangement. In contrast to the nuclear stability regime, **there are dozens or more states with some ability to obtain advanced AI technologies, and those states have a strong incentive to acquire them for multifaceted economic and political reasons that are not necessarily related to any destructive capability such technologies bring**. These states might even

[84] Hin-Yan Liu, Kristian Cedervall Lauta, and Matthijs Michiel Maaset, "Governing Boring Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research," *Futures*, Vol. 102, September 2018, p. 8.

[85] Some have even argued that we might need to build an AI doomsday device to demonstrate its destructive capability. See Matt Chessen, "We Need to Build Doomsday AI," Solarpunk Future, November 26, 2023.

find means to obtain advanced AI technology through espionage campaigns that would allow the theft of models or simply by leveraging openly available capabilities.

This variety of actors extends beyond the number of nuclear weapon states not only because of the nature of the technology but also because of the developing "neomedieval" world in which more states and other actors play major roles in international geopolitics.[86] As a result, **any actual governance regime would likely need to accommodate a larger set of state and other actors than those that were eventually permitted nuclear weapons under the NPT.**

In addition to a larger number of states, **the role of private firms is a major difference from the nuclear stability regime**. Nuclear weapons were a state-driven project funded by governments that maintained possession of the technology. By contrast, **private firms are the primary developers of advanced AI capabilities**, and these companies include some of the largest ones operating worldwide. Competition among these firms has already motivated them to develop and release AI more quickly.[87] Each firm has its own business models and approach to developing and implementing AI, which further increases the complexity of the AI ecosystem to which governance proposals are addressed. AI firms are also proficient lobbyers for creating favorable regulations and shaping public discourse.[88]

Proposals for international AI governance often involve requiring AI developers to be more transparent about model capabilities and risks.[89] Other proposals might also restrict various types of AI development and deployment if they pose significant risks.[90] But **if a company develops a capability that it assesses might have catastrophic impact, it is not entirely clear what that company would be inclined to do**. Depending on the nature of the governance regime, the firm might be required to hand over a trained model to a set of states, an international body, or some nonstate-led body. An AI firm that has a strong conviction might determine that it does not trust the international community or specific governments to keep the capability safe or offer it as a collective benefit for humanity and, thus, might resist efforts to relinquish its capabilities, even leveraging advanced AI as a kind of deterrent against state action.

Thus, governing AI necessitates solving a two-level game, solving for not only strategic interaction among states but also among private-sector firms, which increases the complexity of developing an AI governance regime.

The question of who governs has been answered in the nuclear realm: The NPT lays out the responsibilities of nuclear and nonnuclear nations, entrusts significant governance responsibility to the IAEA, and grants significant informal governance of nuclear proliferation to be exercised by nuclear superpowers, such the United States.

[86] See, for example, Timothy R. Heath, Weilong Kong, and Alexis Dale-Huang, *U.S.-China Rivalry in a Neomedieval World: Security in an Age of Weakening States*, RAND Corporation, RR-A1887-1, 2023.

[87] See, for example, Nilay Patel, "Microsoft Thinks AI Can Beat Google at Search—CEO Satya Nadella Explains Why," The Verge, February 7, 2023.

[88] See, for example, Will Henshall, "There's an AI Lobbying Frenzy in Washington. Big Tech Is Dominating," *Time*, April 30, 2024; and Billy Perrigo, "Exclusive: OpenAI Lobbied the E.U. to Water Down AI Regulation," *Time*, June 20, 2023.

[89] For instance, Executive Order 14110 offers several reporting requirements.

[90] See, for instance, Holden Karnofsky, *If-Then Commitments for AI Risk Reduction*, Carnegie Endowment for International Peace, September 13, 2024.

But it is worth noting that the question of who governs was answered in the 1960s by collusion between the United States and USSR; without their accord, the NPT could not have come into existence, and much of the NPT's negotiation occurred on a bilateral basis between the two superpowers. Once the two nations reached an accommodation, other existing nuclear powers were offered the opportunity to formally join the nuclear club, while other nations that did not yet have nuclear weapons were incentivized to accept their status as nonnuclear nations in exchange for access to civilian nuclear technology. This combination of sticks and carrots provided the underlying incentives for cooperation regarding nuclear nonproliferation.

This historical collusion suggests that, to create a global governance regime for AI, **the global AI powers will need to first identify the mutually agreeable terms for such an agreement**. AI development is concentrated in a few nations, making such a model theoretically possible. However, AI is rapidly proliferating and capable of proliferating much faster than nuclear weapons, suggesting that **the same period of nuclear codominance that characterized the negotiations over the NPT might not hold for AI**. This issue is worsened by the corporate-dominated AI landscape, in which private corporations dominate the development and deployment of AI and are incentivized to rapidly deploy new AI products to reach new customers and withhold technical details because of their proprietary interests. These actors could rapidly proliferate AI to a wide variety of actors, making it difficult to reach the same sort of consensus among a small set of players that characterized the success of the NPT.

In addition, to take a lesson from the Baruch Plan episode, we emphasize the important parallels between the U.S. proposals for the international control of nuclear energy in the late 1940s and contemporary schemes to control AI by depriving rivals of computational resources because these proposals cannot be expected to look fair or equitable to the other side. Just as the Soviet Union had no interest in acceding to the Baruch Plan because, as discussed previously, that plan seemed designed to condemn the USSR to permanent second-class citizenship in nuclear technology, **China might not cooperate with schemes that appear designed with the intention of creating Chinese technological and military inferiority. This is particularly true considering China's focus on reaching technological parity and independence from the United States.**

The example of the Baruch Plan also hints at how badly a policy designed to permanently disadvantage a strategic rival in an important new technology could backfire for the United States. Much as the ostensible U.S. leads in nuclear technology and uranium resources proved to be far smaller and less significant than U.S. officials assumed they were in 1946, the existing U.S. advantages in semiconductor fabrication and compute availability might prove to be short-lived. For example, China might develop a domestic counterpart for Western semiconductor fabrication equipment, and algorithmic advancements could reduce the amount of compute required to train or deploy AI models.[91] **Even if China were called to work with the United States to manage AI risks on a more equitable basis, the bitter seeds of distrust sown by the pursuit of a policy designed to disadvantage China might continue to be an obstacle to cooperation.**

[91] In January 2023, the CEO of Dutch firm ASML, the world's sole manufacturer of extreme ultraviolet lithography machines, remarked of the Chinese, "If they cannot get those machines, they will develop them themselves. That will take time, but ultimately they will get there" (Cagan Koc, "ASML Says Chip Controls Will Push China to Create Own Technology," Bloomberg, January 25, 2023).

Uncertainty About Governance Mechanisms

What form of AI governance could mitigate the high-end catastrophic risks of AI? Of course, an answer to this question depends on the ways that AI might enable catastrophe, and thus, disagreements about risk pathways will carry over to how AI might be governed. But even if there is a clear sense of the ways that AI poses risks, there are difficulties in determining how to mitigate them.

Contemporary AI is sometimes analyzed as composed of three major elements: data, algorithms, and compute.[92] Proposals to govern AI tend to consider how regulations and norms might apply to each element of the AI triad. But **there is disagreement and unclarity about how governance might apply to AI and its various inputs, how it could be monitored and enforced, and what this might mean for reducing AI risk**.

The most advanced AI capabilities are trained on massive quantities of data, and so one might propose that a regulatory approach could restrict the data that a developer uses in AI training to prevent use of data that might provide dangerous capabilities to an AI. However, significant amounts of data are available to be harvested and already well indexed and regularly used for training, even when companies' own rules counsel against it.[93] It might be difficult to **determine which data sources present sufficient marginal risk that it is necessary to restrict their usage in training**, and even if such determinations are made, it might be difficult to control such data sufficiently to prevent their usage for model training. Furthermore, even if potentially dangerous data are restricted, **they might have already been used to train models before the restrictions go into effect, requiring controls on those models that might present additional difficulties to implement.**[94]

The second part of the AI triad, the algorithms or model architecture, might also be an area for governance. If certain architectures seem particularly likely to lead to catastrophic risk—perhaps the transformer models familiar from large language models (LLMs)—one might envision proposing constraints on who could use them and how they could be used. However, similarly to data, **many sophisticated models are either open-source or otherwise possibly taken through espionage.**[95] These models are non-rivalrous, meaning that even as some actors use them, such usage does not prevent others from using them. Ultimately, these algorithms are complex software that can be transferred or obscured from view. Moreover, it is hard to see how computer engineers would be deterred from developing new architectures that might pose increased risk.

[92] See, for example, Ben Buchanan, "The AI Triad and What It Means for National Security Strategy," Center for Security and Emerging Technology, Georgetown University, August 2020.

[93] Cade Metz, Cecilia Kang, Sheera Frenkel, Stuart A. Thompson, and Nico Grant, "How Tech Giants Cut Corners to Harvest Data for A.I.," *New York Times*, April 6, 2024.

[94] It is possible for models to unlearn certain information (see Ronen Eldan and Mark Russinovich, "Who's Harry Potter? Approximate Unlearning in LLMs," arXiv, arXiv:2310.02238, October 4, 2023), although this technique might not be sufficiently dependable to prevent AI from possessing capabilities in dangerous domains.

[95] Leopold Aschenbrenner, "Situational Awareness: The Decade Ahead," webpage, June 2024; Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott, *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Model*, RAND Corporation, RR-A2849-1, 2024.

Given the challenges regulating data and models, analysts have looked to the third part of the triad—compute—as the element that is the most ripe for governance.[96] Indeed, in much writing, compute is seen as the closest analogy to uranium in the nuclear stability regime.[97] Furthermore, the tight supply chain for the semiconductors used to train the most advanced AI might make it easier for intentional bottlenecks that constrain who has access.[98]

To take one recent example, Sastry et al. define four relevant properties of compute that seem to enable governance of AI: detectability, excludability, quantifiability, and supply chain concentration.[99] First, the training and deployment of large-scale AI models is extremely resource-intensive, providing detectability of high-performance clusters that consume significant amounts of power. Second, the physical aspect of the hardware permits users to be excluded from obtaining AI chips, a contrast from data and algorithms, which are intangible and are difficult to control once they are published. Third, the computational power required to develop and deploy AI models can easily be measured, reported, and verified, providing measures of quantifiability. Finally, AI chips are fabricated in a highly complex and inelastic supply chain dominated by a few actors, which potentially enables a limited set of targets for control.[100]

However, there are several known and unknown limitations for compute governance.[101] First, **compute governance is most effective if traceable large-scale training runs continue to be important in producing the most-advanced AI models.**[102] Existing foundational model training runs are expected to cost more than $10 billion in compute, a figure that many expect to keep growing exponentially while LLMs are scaled up.[103] Such runs would involve vast concentrations of chips that would, in turn, be easier to detect and monitor.

[96] Figure 9 in Girish Sastry, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, et al. "Computing Power and the Governance of Artificial Intelligence," arXiv, arXiv:2402.08797, February 13, 2024, is an example of how compute is considered by some analysts to be particularly governable in comparison to other inputs.

[97] See Sastry et al., 2024, Appendix A, for a more detailed analysis of the compute-uranium analogy.

[98] Other writing on this supply chain includes Chris Miller, *Chip War: The Fight for the World's Most Critical Technology*, Scribner, 2022, and Lennart Heim, Markus Anderljung, Emma Bluemke, and Robert Trager, "Computing Power and the Governance of AI," Centre for the Governance of AI, February 14, 2024a.

[99] Sastry et al., 2024.

[100] In a similar schema, Shavit (2023) articulates a verification structure centered on chip inspections and compute monitoring across three levels: on the chip, at the data center, and in the supply chain. Another institutional model centered on compute governance incorporates a multilateral export control procedure that is based on the Nuclear Suppliers Group. This body would be tasked with controlling key inputs for AI systems, such as advanced semiconductors, to ensure that they are only accessible to approved actors. Members of a Compute Suppliers Group would agree to follow guidelines for responsible supplier behavior according to safeguards for AI-related exports and to exchange relevant information with fellow member-states. See Sujai Shivakumar, Charles Wessner, and Hideki Tomoshige, "Toward a New Multilateral Export Control Regime," Center for Strategic and International Studies, January 10, 2023.

[101] In addition to the other limitations of compute governance that we have discussed, Vermeer (2024) describes other key disanalogies, such as the radiation signature of nuclear material. See also Konstantin Pilz, Lennart Heim, and Nicholas Brown, "Increased Compute Efficiency and the Diffusion of AI Capabilities," arXiv, arXiv:2311.15377, February 3, 2024, for a discussion of how compute efficiency might affect the diffusion of AI capabilities and the dangerous implications of particularly capable models.

[102] Sastry et al., 2024.

[103] Aschenbrenner, 2024.

However, there are no guarantees that this factor will continue, despite hypotheses for exponential scaling laws.[104] In fact, there is already evidence that small-scale LLMs focused on a narrow domain could achieve high levels of performance, meaning that risky models could be developed and deployed on relatively smaller amounts of compute, which would be more difficult to track.[105] An illustrative wrinkle is that reasoning models, which reached the market and are increasingly deployed, as of spring 2025, also demonstrate significantly increased capabilities by focusing on inference, enabling models to reach higher levels of capability and have less reliance on training compute or leveraging available pretrained models.[106] In addition, there might emerge alternatives to the common training run paradigms, such as methods to efficiently train LLMs across geographically distributed chips, that might make it difficult to identify when these training runs are happening and how they might be curtailed.[107] The extent to which these issues undermine compute governance is, as of this writing, difficult to predict and might depend on whether compute continues to be deployed in large clusters and accessed through the cloud, which might make compute governance more effective, or whether compute becomes more distributed, which might undermine some compute governance strategies.[108] Although none of these points suggests that compute governance is not useful, they point to wrinkles that might make it more difficult or allow for dangerous capabilities to emerge in models that have relatively less concentration of computing resources, which, in turn, might disperse compute resources further and make effective governance of them more difficult.

Second, **implementing compute governance at a particular point might not account for the hundreds of thousands of machine learning (ML) chips that were acquired previously, which might not be governed by a compute governance program and, therefore, be incapable of being controlled by government**.[109] Although these chips might be less effective than the most-advanced chips because of continued advanced in semiconductors, those chips could still be usable for risky behavior, particularly if increasing compute efficiency also makes it possible to train AI to have risky capabilities using less compute.

In addition, if states perceive significant benefits in developing AI chips that are not subject to compute governance, such states will work to develop domestic compute infrastructure (e.g., China's efforts). One perpetual concern is **whether Western governments can create a regime that adequately constrains China and incentivizes its responsible behavior, largely because of the increased tensions between China and the West**. Although China might be behind the West, China

[104] Fernando Diaz and Michael Madaio, "Scaling Laws Do Not Scale," arXiv, arXiv:2307.03201, July 5, 2023.

[105] Sally Beatty, "Tiny but Mighty: The Phi-3 Small Language Models with Big Potential," Microsoft, April 23, 2024; Tom Taulli, "Small Language Models Gaining Ground at Enterprises," AI Business, January 23, 2024. See also Pilz, Heim, and Brown (2024) for a broader discussion of this topic.

[106] Carter C. Price and Brien Alkire, "What DeepSeek Means for AI Competition: The Beginning of the End or the End of the Beginning," RAND Corporation, February 24, 2025.

[107] For instance, Prime Intellect released a 10-billion parameter model that was "collaboratively trained across the globe" (Johannes Hageman, Sami Jaghouar, Jack Min Ong, and Vincent Weisser, "INTELLECT-1 Release: The First Globally Trained 10B Parameter Model," Prime Intellect, November 29, 2024).

[108] See Lennart Heim, Tim Fist, Janet Egan, Sihao Huang, Stephen Zekany, Robert Trager, Michael A. Osborne, and Noa Zilberman, *Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation*, University of Oxford, AI Governance Initiative, March 13, 2024, for an example of how governance of AI could flow through governance of cloud computing.

[109] Shavit, 2023.

might be able to acquire or develop enough computing power to pose catastrophic threats, despite the imposition of controls on compute access.[110] The success of DeepSeek in developing advanced reasoning models demonstrates that such limitations, even if properly enforced, might not be sufficient to prevent a competing nation from achieving a given capability level, even if the United States is able to maintain an advantage by controlling most compute resources.[111] Finally, similar to other proposals, **these institutional models might involve unprecedented government oversight over the AI industry and private corporations in an industry that has not been subject to such comprehensive controls**, which might hinder widescale buy-in to establish and enforce such a regime.[112]

The question of how to govern involves the complex incentives of states and the potential unanticipated or undesired implications of governance regimes across AI elements. For instance, these regimes might set back economic innovation or other dynamism, especially for those states that do not have access to advanced AI. Any governance regime with haves and have-nots might also be politically unpalatable for states that are deemed untrustworthy because of catastrophic capabilities, so such a regime would need to be accompanied with compelling carrots and sticks.

Thie history of maintaining nuclear stability includes multiple episodes of messy and risky interventions that were undertaken to prevent proliferation beyond simply offering benefits in exchange for compliance. Some AI governance proposals present specific benefits, including the Compute Suppliers Group, the European Organization for Nuclear Research (CERN) for AI, and AI for Peace (see the appendix for a more complete description). But these benefits will also need to be complemented by specific enforcement measures for states that do not comply with international demands. This need for an enforcement mechanism raises the question of what efforts the United States or other countries would undertake to ensure compliance with global AI governance; for instance, would they be willing to engage in expensive sanctions or even more extreme actions that are beyond the scope of this analysis, such as those involving the use of force?

A final challenge has to do with timing: how would this governance infrastructure and international action be established in relatively short order? The NPT entered into force 25 years after Hiroshima and Nagasaki. AI is advancing rapidly, and some analysts suggest that dangerous AI capabilities in such areas as biological weapon creation could emerge within the next few years, as of this writing.[113] There might, therefore, be significantly more urgency to develop an effective AI global governance regime and less time for favorable intellectual and geopolitical developments that might make such a regime more likely to succeed compared with the creation of the NPT.

[110] Meaghan Tobin and Cade Metz, "China Is Closing the A.I. Gap with the United States," *New York Times*, July 25, 2024.

[111] Price and Alkire, 2025. However, controls on chips can still impact AI development in China and beyond, even if it cannot completely prevent model development. See Lennart Heim, "The Rise of DeepSeek: What the Headlines Miss," RAND Corporation, January 28, 2025.

[112] Emma Klein and Stewart Patrick, *Envisioning a Global Regime Complex to Govern Artificial Intelligence*, Carnegie Endowment for International Peace, March 21, 2024.

[113] Bill Drexel and Caleb Withers, *AI and the Evolution of Biological National Security Risks: Capabilities, Thresholds, and Interventions*, Center for a New American Security, August 13, 2024

Conclusion

Given the uncertainties noted previously, a comprehensive global governance agreement for AI will be difficult to achieve. **A primary challenge is the persistent disagreement and uncertainty about the extent of risks posed by existing or future AI technologies, which complicates consensus-building.** Although some technologists and policymakers highlight catastrophic risks, such as misuse of AI to develop destructive technologies or AI-driven disinformation and surveillance, others are skeptical of these risks' severity or propose that the rapid proliferation of AI is, in fact, necessary to achieve socially beneficial goals, such as economic growth. This variety of views and the accompanying mixed set of interests will make it difficult to reach agreement on global governance.

Looking back to the history of nuclear governance reinforces the conclusion that, as of this writing, achieving a global regulatory regime similar to that which exists for nuclear weapons will be very difficult. The successful negotiation of the NPT and its continued viability have rested on a shared interest in nuclear stability and limiting the emergence of new nuclear powers to maintain the comparatively stable status quo among the major nuclear-armed powers. Nuclear nonproliferation was systematized after nearly two decades of nuclear development, during which both the United States and the USSR came to see nonproliferation as in their interest to reduce the threat of proliferation and support nuclear stability between them both. Other states, including many would-be proliferators, have cooperated with the NPT regime because they have concluded that a more familiar world dominated by a few established nuclear powers is preferable to the risks of greater multipolar nuclear competition.

As of this writing, there is no similar conception of AI stability that might incentivize the dominant powers in AI to collude to restrict AI for their benefit. This is not simply a lack of sufficient theorizing regarding why states might collude to institute global AI governance: AI technology is also too immature and its potential too uncertain for firm conclusions on how AI might be used to advance national power or threaten the interests of other countries. Nuclear stability was the result of both a maturing nuclear weapon technology that allowed nuclear powers to hold each other at risk through second-strike capability and a shared understanding of the implications of that technology. In addition, the superpowers recognized the risks of proliferation to nonnuclear states. There is not yet equivalently mature AI technology that has concrete strategic implications nor widely accepted theoretical foundations for understanding AI that might incentivize arms control–like efforts. In fact, AI might never reach a similar point of theoretical and technical maturity because AI has significantly more potential applications and space for technical innovation that might make it difficult to develop a similarly comprehensive theory of AI stability.

Another potentially difficult lesson, arising from the Baruch Plan, is that **there might be difficult trade-offs between strategies wherein a sovereign state seeks to control and exclude others from a particular technology and building a binding governance regime for that technology**. AI is led by the United States, which is the home of the most-advanced AI developers and the foundational intellectual property for the critical semiconductors necessary to economically produce the most-powerful AI models. However, similarly to the Soviet Union, which had no interest in acceding to agreements that seemed designed to condemn the country to permanent second-class citizenship in nuclear technology, contemporary powers, such as China, seem unlikely to cooperate with schemes that appear designed to force them to accept inferiority in AI. This is not to say that strategies to deny access to advanced AI to rivals cannot work but that the experience of the Baruch Plan suggests that the AI rivals of the United States are unlikely to agree to arrangements that they perceive as attempts to institutionalize technological inferiority to the United States.

This is all not to say that international AI governance is impossible, but rather that **the conditions of AI look far more like the chaotic and uncertain period shortly after nuclear weapons were invented rather than the later time when international nuclear weapon governance and nonproliferation were institutionalized**. It took nearly two decades, from the Baruch Plan in 1947 to the successful negotiation of the NPT in 1968, for nuclear governance to succeed. Advocates of AI governance, particularly of comprehensive plans for global regulation, might, therefore, encounter a similarly long and unpredictable journey in advocating for global regulation of AI, despite the potential rapid emergence of AI risks.

However, there might be promise in negotiating the governance of AI in specific policy areas, particularly those that have existing global governance structures and norms that can integrate concerns about AI. Rather than pursuing a comprehensive approach to the risks of AI through new governance regimes modeled after those developed during the Cold War, policymakers might find more traction pursuing governance on narrow pieces of the AI risk landscape, such as the intersection of nuclear stability and AI and, in particular, the ways that AI might increase nuclear instability.

For instance, there are some arguments that AI increases the risk of nuclear war.[114] The existing nuclear stability regime, including such institutions as the IAEA and the set of bilateral and multilateral confidence-building measures, might need to be bolstered or otherwise updated to guard against these new risks, which might involve further research on how AI affects the international strategic balance, how to develop the technical expertise at such institutions as the IAEA, and how to conduct further collaboration between nuclear security experts and private-sector AI developers. It might be helpful to look to history to help bolster these existing regimes, but it will also be necessary to be humble and clear-eyed about the limits of this analysis.

[114] See, for example, Edward Geist and Andrew J. Lohn, *How Might Artificial Intelligence Affect the Risk of Nuclear War?* RAND Corporation, PE-296-RC, April 2018.

Appendix

Proposed Frameworks for International Governance of AI

To categorize the variety of proposals for AI governance, we borrow from the Maas and Villalobos framework, which distinguishes those that create new international institutions to regulate AI according to past and existing institutions from those that create entirely new international institutional models for AI.[115] In their taxonomy of institutional models, Maas and Villalobos also identify a series of distinct functions that have been promoted by scholars and practitioners. For our purposes, we have focused on coordination of policy and regulation (e.g., the World Trade Organization), enforcement of standards or restrictions (e.g., the Nuclear Suppliers Group), stabilization and emergency response (e.g., IAEA), international joint research (e.g., CERN), and distribution of benefits or access (e.g., the IAEA nuclear fuel bank).

International Atomic Energy Agency–Based Institutional Models

The IAEA—more explicitly, its Department of Safeguards—is perceived as having long-standing success at minimizing threats of nuclear war while concurrently overseeing and distributing access to nuclear materials and technologies for nonweapon use. Advocacy for an IAEA-like governing body for AI has been remarkably popular, drawing proponents from AI labs, such as Sam Altman; from international agencies, such as UN Secretary-General António Guterres; and from numerous other researcher and policy institutions.

The IAEA appears in several of the institutional models outlined by Maas and Villalobos. The first approach involves their model for coordination of policy and regulation, which Maas and Villalobos describe as containing an array of functional capabilities that vary among direct regulation, state-assisted implementation of AI policies, harmonization and coordination of policies, certification of industries or jurisdictions, and the monitoring and enforcement of compliance.[116] Altman, Brockman, and Sutskever propose an IAEA for superintelligence that maintains international authority to inspect systems, perform audits, test for safety compliance, and restrict degrees of deployment through levels of security.[117]

Trager et al. propose a version of this model that is loosely based on a combination of the IAEA and CERN, except that it places the responsibility of standards compliance on domestic regulators

[115] Matthijs M. Maas and José Jaime Villalobos, *International AI Institutions: A Literature Review of Models, Examples, and Proposals*, Institute for Law & AI, AI Foundations Report 1, September 2023.

[116] Maas and Villalobos, 2023.

[117] Altman, Brockman, and Sutskever, 2023.

rather than an international organization.[118] The governing body, which the authors nominally call the International AI Organization, would certify jurisdictions' compliance with international standards, which would be developed through a consortium of various stakeholders and enforced through conditional market access to AI technologies. Crucially, this conception relies first on the establishment of clear minimum international regulatory standards followed by states creating their own domestic regulatory capacities for AI that are based on those standards. The authors contend that this model includes several advantages that allow for agile standard-setting, monitoring, and enforcement and could enable rapid responses to standards violations in local jurisdictions.

Other institutional models are those geared toward the enforcement of standards or restrictions, whose function is to "prevent the production, proliferation, or irresponsible deployment of a dangerous or illegal technology, product or activity."[119] Drawing heavily on parallels between uranium material and computer chips, Baker offers a case study for hardware-based monitoring and verification mechanisms for AI that is reflective of the IAEA model. Under this regime, computer chips used for compute-intensive AI development are required to contain a built-in mechanism to enable verification. Importantly, Baker argues that preliminary preparations to develop private, secure, and cost-effective methods of verification and the establishment of an incomplete but easy-to-improve verification system can help prevent some foreseeable challenges of AI treaty verification that were successfully abated in nuclear arms control.[120]

In a similar schema, Shavit articulates further details of a verification structure centered on chip inspections and compute monitoring across three levels: on the chip, at the data center, and in the supply chain. To prove compliance, the owners of ML chips would employ firmware to log information about the chip's activity. Inspectors can then observe the logs of a sufficient sample of ML chips to determine whether the chip owner violated established rules for a training run. This institutional model has two important limitations. First, it requires that traceable large-scale training runs will continue to be important in producing the most-advanced AI models, thus maintaining frontier models' dependence on advanced chips. Second, it does not account for the hundreds of thousands of ML chips acquired previously, which do not contain the hardware security features necessary for the framework and might be incapable of being retrofitted or located by governments.[121]

The creation of an AI governance entity has emerged as a proposed solution to harmonize international oversight and regulation of AI systems. Gutierrez explores adjacent avenues for an institutional model of AI governance that combines the UN's authority to create specialized agencies, such as the IAEA, with the Intergovernmental Panel on Climate Change and Interpol's alert system.[122] The proposal is structured around three pillars for the proposed AI governance entity's work: identifying paths of shared large-scale high-risk harms from AI systems; coordinating

[118] Robert Trager, Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, Ranjit Lall, Owen Larter, Seán Ó hÉigeartaigh, et al. "International Governance of Civilian AI: A Jurisdictional Certification Approach," arXiv, arXiv:2308.15514, August 29, 2023.

[119] Maas and Villalobos, 2023, p. 24.

[120] Baker, 2023.

[121] Shavit, 2023.

[122] Carlos I. Gutierrez, "Multilateral Coordination for the Proactive Governance of Artificial Intelligence Systems," Future of Life Institute, September 25, 2023.

technically sound global responses that is consistent with governance best practices; and enforcing commitments to actions through mutual agreement on the ability to reduce the likelihood and severity of harms.

Compute Governance

As identified in Shavit's proposed international model, the governance of compute—a quantifiable measure of the computational power required to train AI models and perform task—might provide leverage to contain high-risk scenarios from AI development.[123] Efforts to define compute thresholds are already underway in the United States and Europe. For example, President Biden's now-revoked Executive Order 14110 on AI issued in October 2023 and the EU AI Act include transparency and other requirements for models trained over specific computing thresholds.[124] Compute governance might also be carried out through such mechanisms as export controls, which the United States and its partners have used to restrict access to the advanced chips used in AI to competitor nations. These mechanisms control the supply of compute rather than implementing governance mechanisms directly on chips, as proposed by Shavit, and accomplish a slightly different end of restricting access to compute for certain actors rather than ensuring all compute be equipped with governance mechanisms.

Sastry et al. define four properties of compute that enable the governance of AI: detectability, excludability, quantifiability, and supply chain concentration.[125] First, the training and deployment of large-scale AI models continues to be extremely resource-intensive, providing detectability of high-performance clusters that consume significant amounts of power. Second, the physical aspect of the hardware permits users to be excluded from obtaining AI chips, a contrast from data and algorithms, which are intangible and are difficult to control once they are published. Third, the computational power required to develop and deploy AI models can easily be measured, reported, and verified, providing measures of quantifiability. Finally, because AI chips are fabricated in a highly complex and inelastic supply chain, the crucial foundational steps in AI development are dominated by a small number of actors, making it easier to oversee.[126]

Compute Suppliers Group

Extending beyond the establishment of a new or revised UN agency to oversee AI advancements and address risks, some proposals have incorporated a multilateral export control procedure akin to the Nuclear Suppliers Group. This institutional model would be tasked with controlling key inputs for AI systems, such as advanced semiconductors, to ensure that they are only accessible to approved

[123] Shavit, 2023.

[124] See Executive Order 14110, 2023; and European Commission, "General-Purpose AI Models in the AI Act – Questions & Answers," webpage, March 14, 2025, for the thresholds used to categorize general-purpose AI models according to compute used.

[125] Sastry et al., 2024.

[126] Sastry et al., 2024.

actors.[127] Members of a Compute Suppliers Group would agree to follow guidelines for responsible supplier behavior according to safeguards for AI-related exports and to exchange relevant information with fellow member-states.

Klein and Patrick identify a few noteworthy limitations that resemble those of other institutional models. One central concern is whether Western governments can create a regime that adequately constrains China and incentivizes its responsible behavior, largely because of the aforementioned increased tensions between China and the West. Another potential pitfall is the familiar intangible nature of AI models and algorithms, which makes it difficult to enforce limits on digital inputs outside chips and hardware. Finally, similar to other proposals, this model would require unprecedented government oversight over the AI industry and private corporations, which might be untenable with major players and hinder widescale buy-in.[128]

European Organization for Nuclear Research and International Joint Research Agreements

CERN is another prominent international body that draws comparisons to AI in institutional models for governance that Maas and Villalobos identify as international joint research agreements.[129] These function through a bilateral or multilateral partnership between national states or state entities to collaborate on solving common scientific problems or achieving a common goal. Sastry et al. also provides insights into this institutional model and identifies several technical objectives for a CERN for AI.[130] More generally, a CERN for AI could provide computing resources to AI labs conducting large research projects or could focus on safely and equitably training frontier models for broad societal benefit. Another objective could be to concentrate on public goods, such as clean energy, sustainability, and the medical research applications of AI. The institution has the potential to encourage cooperation between competing countries by building trust and stabilizing future AI arms races.

Non-Proliferation Treaty–Based Proposals

The NPT is a historical arrangement that has been explored for lessons for AI governance. Central to the NPT is the "core bargain, whereby non-nuclear-weapons states agree not to acquire such weapons in return for a pledge by the five acknowledged nuclear-weapons states to pursue nuclear disarmament and share the benefits of access to peaceful nuclear technology."[131] In terms of AI, Article IV of the NPT is the most relevant and authorizes access to peaceful applications of nuclear technology on the condition that member states abstain from pursuing nuclear weapons and congruently agree to consider the needs of the Global South and developing areas of the world in the

[127] Shivakumar, Wessner, and Tomoshige, 2023.

[128] Klein and Patrick, 2024.

[129] Maas and Villalobos, 2023.

[130] Sastry et al., 2024.

[131] Klein and Patrick, 2024, p. 21.

process. Therefore, proposals for AI would commit all treaty members to ensure that access to equipment, materials, and scientific and technological information is distributed to all member nations.

AI for Peace

Dwight Eisenhower's "Atoms for Peace" speech is widely considered to be a foundational precursor to the IAEA and nuclear nonproliferation agreements and has also been a source for analogies between AI and nuclear technologies.[132] Roberts lays out a review of these plans centered on a variety of desirable norms to be encouraged through an AI for Peace model. These range from a no-kill rule for AI that requires a human-in-the-loop to be responsible for military attacks to the addition of an off switch to shut down an AI system for maintenance or when the AI system might pose a threat. The core responsibilities of this model would be to clearly convey the dangers of AI systems, construct principles to alleviate the risks from the dangers, mediate access to the resources needed to develop and deploy AI systems, and shape the incentives for states through a system of monitoring and inspection.[133]

[132] Eisenhower, 1953.

[133] Patrick S. Roberts, "AI for Peace," *War on the Rocks*, December 13, 2019.

Abbreviations

AI	artificial intelligence
CERN	European Organization for Nuclear Research
IAEA	International Atomic Energy Agency
LLM	large language model
MAD	mutually assured destruction
ML	machine learning
MLF	multilateral force
NATO	North Atlantic Treaty Organization
NPT	Treaty on the Non-Proliferation of Nuclear Weapons, commonly known as the Non-Proliferation Treaty
UN	United Nations
UNAEC	United Nations Atomic Energy Commission
USSR	Union of Soviet Socialist Republics

References

Unless otherwise indicated, the authors of this paper provided the translations of bibliographic details for the non-English sources included in this paper. To support conventions for alphabetizing, the source in Russian is introduced with and organized according to its English translation. The original rendering in Russian appears in brackets after the English translation.

Altman, Sam, Greg Brockman, and Ilya Sutskever, "Governance of Superintelligence," OpenAI blog, May 22, 2023.

Amodei, Dario, "Oversight of A.I.: Principles for Regulation," testimony before the U.S. Senate Judiciary Committee Subcommittee on Privacy, Technology, and the Law, July 25, 2023.

Anderson, Justin, Jeffrey Larsen, and Polly Holdorf, *Extended Deterrence and Allied Assurance: Key Concepts and Current Challenges for U.S. Policy*, U.S. Air Force Institute for National Security Studies, September 2013.

Aschenbrenner, Leopold, "Situational Awareness: The Decade Ahead," webpage, June 2024. As of May 2, 2025:
https://situational-awareness.ai/

Baker, Mauricio, "Nuclear Arms Control Verification and Lessons for AI Treaties," arXiv, arXiv:2304.04123, April 8, 2023.

The Baruch Plan, presented to the United Nations Atomic Energy Commission, June 14, 1946.

Beatty, Sally, "Tiny but Mighty: The Phi-3 Small Language Models with Big Potential," Microsoft, April 23, 2024.

Blackett, Patrick Maynard Stuart, *Fear, War, and the Bomb: Military and Political Consequences of Atomic Energy*, Whittlesey House, 1949.

Boyer, Paul, *By the Bomb's Early Light: American Thought and Culture at the Dawn of the Atomic Age*, Pantheon, 1985.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," arXiv, arXiv:2303.12712, April 13, 2023.

Buchanan, Ben, "The AI Triad and What It Means for National Security Strategy," Center for Security and Emerging Technology, Georgetown University, August 2020.

Bunn, George, "The Nuclear Nonproliferation Treaty: History and Current Problems," Arms Control Association, December 2003.

Buolamwini, Joy, *Unmasking AI: My Mission to Protect What Is Human in a World of Machines*, Random House, 2023.

Burr, William, ed., "Detection of the First Soviet Nuclear Test, September 1949," National Security Archive, September 9, 2016.

Burr, William, ed., "Preoccupations with West Germany's Nuclear Weapons Potential Shaped Kennedy-Era Diplomacy," National Security Archive, February 2, 2018a.

Burr, William, ed., "The Nuclear Nonproliferation Treaty and the German Nuclear Question Part II, 1965–1969," National Security Archive, March 21, 2018b.

Burr, William, and David Rosenberg, "Nuclear Competition in the Age of Stalemate, 1963–1975," in Melvin P. Leffler and Odd Anne Westad, eds., *The Cambridge History of the Cold War*, Vol. II, *Crises and Détente*, Cambridge University Press, 2010.

Center for AI Safety, "Statement on AI Risk," undated.

Center for Arms Control and Non-Proliferation, "Fact Sheet: Thermonuclear Weapons," November 18, 2022.

Chessen, Matt, "We Need to Build Doomsday AI," Solarpunk Future, November 26, 2023.

Chesterman, Simon, "Weapons of Mass Disruption: Artificial Intelligence and International Law," *Cambridge International Law Journal*, Vol. 10, No. 2, December 2021.

Coe, Andrew J., and Jane Vaynman, "Collusion and the Nuclear Nonproliferation Regime," *Journal of Politics*, Vol. 77, No. 4, January 2015.

Colby, Elbridge A., and Michael S. Gerson, eds., *Strategic Stability: Contending Interpretations*, Strategic Studies Institute and U.S. Army War College Press, February 2013.

Craig, Campbell, and Sergey Radchenko, *The Atomic Bomb and the Origins of the Cold War*, Yale University Press, 2008.

Diaz, Fernando, and Michael Madaio, "Scaling Laws Do Not Scale," arXiv, arXiv:2307.03201, July 5, 2023.

Drexel, Bill, and Caleb Withers, *AI and the Evolution of Biological National Security Risks: Capabilities, Thresholds, and Interventions*, Center for a New American Security, August 13, 2024

Eisenhower, Dwight D., "Atoms for Peace Speech," address to the 470th Plenary Meeting of the United Nations General Assembly, December 8, 1953.

Eldan, Ronen, and Mark Russinovich, "Who's Harry Potter? Approximate Unlearning in LLMs," arXiv, arXiv:2310.02238, October 4, 2023.

European Commission, "General-Purpose AI Models in the AI Act – Questions & Answers," webpage, March 14, 2025.

Executive Order 14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Executive Office of the President, October 30, 2023.

Fairbanks, Charles H., Jr., "MAD and U.S. Strategy," in Henry D. Sokolski, ed., *Getting MAD: Nuclear Mutual Assured Destruction, Its Origin and Practice*, U.S. Army War College Press, 2004.

Fulham, Mary Olney, "Ask the Experts: The 60th Anniversary of the Cuban Missile Crisis," Nuclear Threat Initiative, October 13, 2022.

Gavin, Francis J., "Strategies of Inhibition: U.S. Grand Strategy, the Nuclear Revolution, and Non-Proliferation," *International Security*, Vol. 40, No. 1, Summer 2015, p. 21.

Geist, Edward, and Andrew J. Lohn, *How Might Artificial Intelligence Affect the Risk of Nuclear War?* RAND Corporation, PE-296-RC, April 2018. As of April 23, 2025: https://www.rand.org/pubs/perspectives/PE296.html

Goldschmidt, Bertrand, *A Forerunner of the NPT? The Soviet Proposals of 1947*, International Atomic Energy Agency, Vol. 28-1, March 1986.

Gordin, Michael D., *Red Cloud at Dawn: Truman, Stalin, and the End of the Atomic Monopoly*, Farrar, Straus and Giroux, 2009.

Gutierrez, Carlos I., "Multilateral Coordination for the Proactive Governance of Artificial Intelligence Systems," Future of Life Institute, September 25, 2023.

Hageman, Johannes, Sami Jaghouar, Jack Min Ong, and Vincent Weisser, "INTELLECT-1 Release: The First Globally Trained 10B Parameter Model," Prime Intellect, November 29, 2024.

Heath, Timothy R., Weilong Kong, and Alexis Dale-Huang, *U.S.-China Rivalry in a Neomedieval World: Security in an Age of Weakening States*, RAND Corporation, RR-A1887-1, 2023. As of April 23, 2025: https://www.rand.org/pubs/research_reports/RRA1887-1.html

Heim, Lennart, "The Rise of DeepSeek: What the Headlines Miss," RAND Corporation, January 28, 2025. As of April 23, 2025: https://www.rand.org/pubs/commentary/2025/01/the-rise-of-deepseek-what-the-headlines-miss.html

Heim, Lennart, Markus Anderljung, Emma Bluemke, and Robert Trager, "Computing Power and the Governance of AI," Centre for the Governance of AI, February 14, 2024.

Heim, Lennart, Tim Fist, Janet Egan, Sihao Huang, Stephen Zekany, Robert Trager, Michael A. Osborne, and Noa Zilberman, *Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation*, University of Oxford, AI Governance Initiative, March 13, 2024.

Henshall, Will, "There's an AI Lobbying Frenzy in Washington. Big Tech Is Dominating," *Time*, April 30, 2024.

Herken, Gregg, *The Winning Weapon: The Atomic Bomb in the Cold War, 1945–1950*, Alfred A. Knopf, 1980.

Holloway, David, *Stalin and the Bomb: The Soviet Union and Atomic Energy, 1939–1956*, Yale University Press, 1994.

Hunt, Johathan, "'If One Tightens the Screw to the Limit . . . One Might Strip the Thread': Soviet Defenses of the Nuclear Non-Proliferation Treaty," Sources and Methods blog, September 7, 2023. As of April 22, 2025: https://www.wilsoncenter.org/blog-post/if-one-tightens-screw-limit-one-might-strip-thread-soviet-defenses-nuclear-non

Kahn, Herman, *On Thermonuclear War*, Princeton University Press, 1960.

Karnofsky, Holden, *If-Then Commitments for AI Risk Reduction*, Carnegie Endowment for International Peace, September 13, 2024.

Kaushik, Divyansh, and Matt Korda, "Panic About Overhyped AI Risk Could Lead to the Wrong Kind of Regulation," Vox, July 3, 2023.

Kemp, R. Scott, "The End of Manhattan: How the Gas Centrifuge Changed the Quest for Nuclear Weapons," *Technology and Culture*, Vol. 53, No. 2, April 2012.

Klein, Emma, and Stewart Patrick, *Envisioning a Global Regime Complex to Govern Artificial Intelligence*, Carnegie Endowment for International Peace, March 21, 2024.

Klyman, Kevin, and Raphael Piliero, "AI and the A-Bomb: What the Analogy Captures and Misses," *Bulletin of the Atomic Scientists*, September 9, 2024.

Koc, Cagan, "ASML Says Chip Controls Will Push China to Create Own Technology," Bloomberg, January 25, 2023.

Krige, John, "The Proliferation Risks of Gas Centrifuge Enrichment at the Dawn of the NPT: Shedding Light on the Negotiating History," *Nonproliferation Review*, Vol. 19, No. 2, July 2012.

Liu, Hin-Yan, Kristian Cedervall Lauta, and Matthijs Michiel Maaset. "Governing Boring Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research," *Futures*, Vol. 102, September 2018.

Lutsch, Andreas, "The Federal Republic of Germany and the NPT, 1967–1969," Sources and Methods blog, January 29, 2024. As of April 22, 2025:
https://www.wilsoncenter.org/blog-post/federal-republic-germany-and-npt-1967-1969

Maas, Matthijs M., "How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons," *Contemporary Security Policy*, Vol. 40, No. 3, February 2019.

Maas, Matthijs M., and José Jaime Villalobos, *International AI Institutions: A Literature Review of Models, Examples, and Proposals*, Institute for Law & AI, AI Foundations Report 1, September 2023.

Matheny, Jason, "Challenges to U.S. National Security and Competitiveness Posed by AI," testimony presented before the U.S. Senate Committee on Homeland Security and Governmental Affairs, RAND Corporation, CT-A2654-1, March 8, 2023. As of April 22, 2025:
https://www.rand.org/pubs/testimonies/CTA2654-1.html

Matthews, Dylan, "AI Is Supposedly the New Nuclear Weapons—but How Similar Are They, Really?" Vox, June 29, 2023.

Mazarr, Michael J., *Understanding Deterrence*, RAND Corporation, PE-295-RC, April 2018. As of April 22, 2025:
https://www.rand.org/pubs/perspectives/PE295.html

Metz, Cade, Cecilia Kang, Sheera Frenkel, Stuart A. Thompson, and Nico Grant, "How Tech Giants Cut Corners to Harvest Data for A.I.," *New York Times*, April 6, 2024.

Miller, Chris, *Chip War: The Fight for the World's Most Critical Technology*, Scribner, 2022.

Mouton, Christopher A., Caleb Lucas, and Ella Guest, *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study*, RAND Corporation, RR-A2977-2, 2024. As of October 4, 2024:
https://www.rand.org/pubs/research_reports/RRA2977-2.html

Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott, *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*, RAND Corporation, RR-A2849-1, 2024. As of October 4, 2024:
https://www.rand.org/pubs/research_reports/RRA2849-1.html

Ord, Toby, *The Precipice: Existential Risk and the Future of Humanity*, Bloomsbury Publishing, 2020.

Patel, Nilay, "Microsoft Thinks AI Can Beat Google at Search—CEO Satya Nadella Explains Why," The Verge, February 7, 2023.

Patwardhan, Tejal, Kevin Liu, Todor Markov, Neil Chowdhury, Dillon Leet, Natalie Cone, Caitlin Maltbie, Joost Huizinga, Carroll Wainwright, Shawn (Froggi) Jackson, et al., "Building an Early Warning System for LLM-Aided Biological Threat Creation," OpenAI, January 31, 2024.

Perrigo, Billy, "Exclusive: OpenAI Lobbied the E.U. to Water Down AI Regulation," *Time*, June 20, 2023.

Pilz, Konstantin, Lennart Heim, and Nicholas Brown, "Increased Compute Efficiency and the Diffusion of AI Capabilities," arXiv, arXiv:2311.15377, February 3, 2024.

Piper, Kelsey, "Is It Time for a Pause?" Planned Obsolescence blog, March 30, 2023.

Popp, Roland, Liviu Horovitz, and Andreas Wagner, eds., *Negotiating the Nuclear Non-Proliferation Treaty: Origins of the Nuclear Order*, Routledge Taylor & Francis Group, 2017.

Price, Carter C., and Brien Alkire, "What DeepSeek Means for AI Competition: The Beginning of the End or the End of the Beginning," RAND Corporation, February 24, 2025. As of April 23, 2025: https://www.rand.org/pubs/commentary/2025/02/what-deepseek-means-for-ai-competition-the-beginning.html

Roberts, Patrick S., "AI for Peace," *War on the Rocks*, December 13, 2019.

Sastry, Girish, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, et al. "Computing Power and the Governance of Artificial Intelligence," arXiv, arXiv:2402.08797, February 13, 2024.

Shavit, Yonadav, "What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training Via Compute Monitoring," arXiv, arXiv:2303.11341, May 30, 2023.

Shivakumar, Sujai, Charles Wessner, and Hideki Tomoshige, "Toward a New Multilateral Export Control Regime," Center for Strategic and International Studies, January 10, 2023.

Solomon, James B., *The Multilateral Force: America's Nuclear Solution for NATO (1960–1965)*, Naval Academy, May 4, 1999.

"Stop Talking About Tomorrow's AI Doomsday When AI Poses Risks Today," *Nature*, Vol. 618, No. 7967, June 2023.

Tarasenko, V. A., *The Atomic Problem in the External Relations of the U.S.A.* [Атомная проблема во внешней политике США], Shevchenko Kyiv State University Press [Izd-vo Kievskogo gosudarstvennogo universiteta im. T. G. Shevchenka], 1958.

Taulli, Tom, "Small Language Models Gaining Ground at Enterprises," AI Business, January 23, 2024.

Tiku, Nitasha, "The Google Engineer Who Thinks the Company's AI Has Come to Life," *Washington Post*, June 11, 2022.

Tobin, Meaghan, and Cade Metz, "China Is Closing the A.I. Gap with the United States," *New York Times*, July 25, 2024.

Trager, Robert, Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, Ranjit Lall, Owen Larter, Seán Ó hÉigeartaigh, et al. "International Governance of Civilian AI: A Jurisdictional Certification Approach," arXiv, arXiv:2308.15514, August 29, 2023.

Treaty on the Non-Proliferation of Nuclear Weapons, signed at London, United Kingdom; Moscow, Russia; and Washington, D.C., on July 1, 1968.

United Nations, "International Community Must Urgently Confront New Reality of Generative, Artificial Intelligence, Speakers Stress as Security Council Debates Risks, Rewards," July 18, 2023.

Vermeer, Michael J. D., *Historical Analogues That Can Inform AI Governance*, RAND Corporation, RR-A3408-1, 2024. As of April 22, 2025:
https://www.rand.org/pubs/research_reports/RRA3408-1.html

Wallace, Henry A., "From the Letter to the President," *Bulletin of the Atomic Scientists*, Vol. 2, Nos. 7–8, 1946.

Watson, Mike, "IAEA for AI? That Model Has Already Failed," *Wall Street Journal*, June 1, 2023.

Weiss, Leonard, "Nuclear-Weapon States and the Grand Bargain," Arms Control Association, December 2003.

Wohlstetter, Albert, Fred Hoffman, R. J. Lutz, and Henry S. Rowen, *Selection and Use of Strategic Air Bases*, RAND Corporation, R-266, 1954. As of July 28, 2024:
https://www.rand.org/pubs/reports/R0266.html

Wong, Matteo, "AI Doomerism Is a Decoy," *The Atlantic*, June 2, 2023.

Xu, Tammy, "We Could Run Out of Data to Train AI Language Programs," *MIT Technology Review*, November 24, 2022.

Zaidi, Waqar, and Allan Dafoe, "International Control of Powerful Technology: Lessons from the Baruch Plan for Nuclear Weapons," Centre for the Governance of AI, March 2021.

About the Authors

Benjamin Boudreaux is a policy researcher at RAND working in the intersection of human security, technology, and ethics. His current research focuses on AI ethics and governance. He holds a Ph.D. in philosophy.

Gregory Smith is a policy analyst at RAND. His research interests include critical and emerging technologies, the impact and governance of powerful artificial intelligence, defense innovation, U.S.-China and Indo-Pacific security issues, international trade and finance, and supply chain security. He holds a J.D.

Edward Geist is a senior policy researcher at RAND. His research interests include the former Soviet Union, nuclear weapons, emergency management, and artificial intelligence. He holds a Ph.D. in Russian history.

Leah Dion is a technical analyst at RAND. Her research interests span the gamut of RAND's policy portfolio, with a special interest in justice policy and emerging technologies as they relate to socioeconomic inequities. She holds an M.S. in data analytics and computational social science.