R Street
Free markets. Real solutions.

R Street Policy Study
No. 291
June 2023

# Existential Risks and Global Governance Issues Around AI and Robotics

By Adam Thierer

**Continuous communication, coordination and cooperation—among countries, developers, professional bodies and other stakeholders—will be essential in heading off risks as they develop and in creating and reinforcing ethical norms.**

## Table of Contents

## Executive Summary

There are growing concerns about how lethal autonomous weapons systems, artificial general intelligence (or “superintelligence”) or “killer robots” might give rise to new global existential risks. Continuous communication and coordination—among countries, developers, professional bodies and other stakeholders—is the most important strategy for addressing such risks.

Although global agreements and accords can help address some malicious uses of artificial intelligence (AI) or robotics, proposals calling for control through a global regulatory authority are both unwise and unlikely to work. Calls for bans or “pauses” on AI developments are also futile because many nations would never agree to forego developing algorithmic capabilities when adversaries are advancing their own. Therefore, the U.S. government should continue to work with other nations to address threatening uses of algorithmic or robotic technologies while simultaneously taking steps to ensure that it possesses the same technological capabilities as adversaries or rogue nonstate actors.

Many different nongovernmental international bodies and multinational actors can play an important role as coordinators of national policies and conveners of ongoing deliberation about various AI risks and concerns. Soft law (i.e., informal rules, norms and agreements) will also play an important role in addressing AI risks. Professional institutions and nongovernmental bodies have developed important ethical norms and expectations about acceptable uses of algorithmic technologies, and these groups also play an essential role in highlighting algorithmic risks and helping with ongoing efforts to communicate and coordinate global steps to address them.

## Introduction: The Realpolitik of Global AI Governance

The so-called "existential risks" surrounding AI and robotics are attracting increasing academic and governmental attention, with headlines warning of how artificial intelligence (AI) and artificial-generated intelligence (AGI)—or "superintelligent" AI—could "defeat all of us combined" or "kill everyone."[1] These are risks that raise the specter of extraordinary threats to life, limb, health, political stability, public order and human survival, and they are garnering more attention as global AI competition intensifies and the potential for malicious uses of computational systems expands.[2]

Analysts and policymakers have different concerns when discussing global AI risks. Some worry that the United States could fall behind other nations. Others worry about algorithmic systems fueling state propaganda and misinformation efforts or runaway/unaligned AI that might undermine human values or public safety.

Analysts and policymakers have different primary concerns when discussing global AI risks. Some worry that the United States could fall behind other nations in terms of technological readiness.[3] In particular, growing concerns about China's technological capabilities have resulted in a flurry of hearings, events and major reports, often driven by fears of an "artificial intelligence Cold War on the horizon" and "the militarization of artificial intelligence" as part of a growing class of data-driven, nonkinetic weapons.[4] Others worry about algorithmic systems fueling state propaganda and misinformation efforts or runaway/unaligned AI that might undermine human values or public safety.[5]

Some analysts argue that dangers such as these represent a new type of global catastrophic risk, prompting calls for a global regulatory body and a set of international laws to address them.[6] Even some major AI developers have raised such concerns. For example, Elon Musk has warned, "[w]ith artificial intelligence, we

[1] Holden Karnofsky, "AI Could Defeat All Of Us Combined," Cold Takes, June 9, 2022. https://www.cold-takes.com/ai-could-defeat-all-of-us-combined; Sarah Knapton, "Advanced AI 'could kill everyone', warn Oxford researchers," *The Telegraph*, Jan. 25, 2023. https://www.telegraph.co.uk/news/2023/01/25/advanced-ai-could-kill-everyone-warn-oxford-researchers.

[2] Toby Ord, *The Precipice: Existential Risk and the Future of Humanity* (Hachette Books, 2020); Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," Future of Humanity Institute, February 2018. https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf.

[3] Eric Schmidt, "Innovation Power: Why Technology Will Define the Future of Geopolitics," *Foreign Affairs*, Feb. 28, 2023. https://www.foreignaffairs.com/united-states/eric-schmidt-innovation-power-technology-geopolitics.

[4] Ryan Heath, "Artificial Intelligence Cold War on the horizon," *Politico*, Oct. 16, 2020. https://www.politico.com/news/2020/10/16/artificial-intelligence-cold-war-on-the-horizon-429714; Paul Scharre, "4. The Militarization of Artificial Intelligence," Texas National Security Review, June 2, 2020. https://tnsr.org/roundtable/policy-roundtable-artificial-intelligence-and-international-security/#essay4.

[5] Bill Drexel and Caleb Withers, "Generative AI could be an authoritarian breakthrough in brainwashing," *The Hill*, Feb. 26, 2023. https://thehill.com/opinion/technology/3871841-generative-ai-could-be-an-authoritarian-breakthrough-in-brainwashing; Sigal Samuel, "Effective altruism's most controversial idea," *Vox*, Sept. 6, 2022. https://www.vox.com/future-perfect/23298870/effective-altruism-longtermism-will-macaskill-future.

[6] Olivia J. Erdélyi and Judy Goldsmith, "Regulating Artificial Intelligence: Proposal for a Global Solution," *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (December 2018), pp. 95-101. https://dl.acm.org/doi/10.1145/3278721.3278731.

are summoning the demon," and Sam Altman, the head of OpenAI and the creator of GPT-4 and ChatGPT suggests that these issues "probably do need a … global regulatory body."[7]

In March 2023, the Future of Life Institute released an open letter that included some notable computer science experts calling for AI labs "to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4."[8] One prominent AI ethicist went further, insisting that "pausing AI developments isn't enough," suggesting that governments should consider the use of airstrikes against data processing centers—potentially even with nuclear weapons—to stop the development of powerful computational systems.[9] Others have proposed placing caps on how large or powerful developers can make large-scale AI models without prior approval from a regulatory authority.[10]

Thus, there are many distinct issues, but no one-size-fits-all solution.[11] Importantly, proposed regulatory solutions could give rise to other risks that may be even more serious, such as curtailing algorithmic innovations, limiting liberties or creating hostilities between nations. The "realpolitik" of international AI governance will therefore necessitate balanced and pragmatic responses.

There are many distinct issues, but no one-size-fits-all solution. Importantly, proposed regulatory solutions could give rise to other risks that may be even more serious, such as curtailing algorithmic innovations, limiting liberties or creating hostilities between nations.

This study considers what sort of governance responses are realistic when looking at various types of potential global AI risks. First, we look to the nature of existential risks in the context of the global stage and argue that academic and political use of the term requires greater precision. With a clearer definition in mind, we then explore how sweeping responses to AI risks—many of which could block future innovation and scientific progress—could also give rise to new existential risks by depriving society of new technologies that may reduce existing risks and help advance public health and safety. The paper next considers existing and proposed frameworks for addressing global algorithmic risks and discusses how these approaches might help address concerns about "killer robots" or various lethal autonomous weapons systems (LAWS).

We conclude with a series of key findings, based on the understanding that proposals hoping to impose global controls through a worldwide regulatory authority are both unwise and unlikely to work. Calls for bans or "pauses" on AI or supercomputing are largely futile because most nations will not agree to them. Additionally, we must identify the problems associated with mass surveillance

[7] Samuel Gibbs, "Elon Musk: artificial intelligence is our biggest existential threat," *The Guardian*, Oct. 27, 2014. https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat; Kara Swisher, "Sam Altman on What Makes Him 'Super Nervous' About AI," *Intelligencer*, March 23, 2023. https://nymag.com/intelligencer/2023/03/on-with-kara-swisher-sam-altman-on-the-ai-revolution.html.

[8] "Pause Giant AI Experiments: An Open Letter," Future of Life Institute, March 22, 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments.

[9] Eliezer Yudkowsky, "Pausing AI Developments Isn't Enough. We Need to Shut it All Down," *Time*, March 29, 2023. https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough.

[10] Kelsey Piper, "A.I. Is About to Get Much Weirder. Here's What to Watch For.," *The New York Times*, March 21, 2023. https://www.nytimes.com/2023/03/21/opinion/ezra-klein-podcast-kelsey-piper.html.

[11] Heather Frase, "One Size Does Not Fit All: Assessment, Safety, and Trust for the Diverse Range of AI Products, Tools, Services, and Resources," Center for Security and Emerging Technology, February 2023. https://cset.georgetown.edu/publication/one-size-does-not-fit-all.

solutions meant to control the flow of information or limit research and development for algorithmic or robotic systems. The realpolitik of global AI policy will demand a variety of alternative coordination and communication efforts aimed at identifying and addressing algorithmic risks in real time. There are no silver-bullet solutions. Many different nongovernmental international bodies, and multinational actors—both governmental and nongovernmental—will need to play a role as coordinators of pragmatic oversight policies and conveners of ongoing dialogue about various AI risks and concerns.

## The Problem with the Precautionary Principle for AI Policy

Previous R Street research has argued that the wisest policy default for AI and robotics continues to be permissionless innovation—or a general freedom to research and develop new algorithmic capabilities—rather than the precautionary principle, which generally restricts innovation until technologies have been approved by a regulatory authority.[12] Another R Street Institute report explained that calls for AI "safety by design" are sensible and outlined how that goal can be achieved in a more flexible, decentralized fashion using a wide variety of agile, bottom-up governance tools and methodologies.[13]

If the precautionary principle were to become the legal standard for AI developers, they would need to "affirmatively demonstrate that their technology is not harmful and self-certify or seek regulatory approval before they deploy it." Such a mandate would greatly limit innovation potential in the AI and robotics fields.

Although the precautionary principle reflects a well-intentioned desire to play it safe in the face of uncertainty, it gives rise to serious problems when translated to regulatory mandates.[14] For example, some regulatory advocates have proposed treating algorithmic innovations under a standard of "unlawfulness by default."[15] If this were to become the legal standard for AI developers, they would need to "affirmatively demonstrate that their technology is not harmful and self-certify or seek regulatory approval before they deploy it."[16] Such a mandate would greatly limit innovation potential in the AI and robotics fields.

This is why some scholars speak of the hidden costs "of saying no" associated with precautionary principle restraints, which include lost products and services; higher prices; diminished economic vitality and growth; fewer employment opportunities; and more.[17] These constraints can also derail the learning curve by limiting opportunities to gain important insights from trial-and-error experimentation with new and better ways of doing things.[18] In fact, historians have documented how,

[12] Adam Thierer, "Getting AI Innovation Culture Right," *R Street Institute Policy Study* No. 281, March 30, 2023. https://www.rstreet.org/research/getting-ai-innovation-culture-right.

[13] Adam Thierer, "Flexible, Pro-Innovation Governance Strategies for Artificial Intelligence," *R Street Institute Policy Study* No. 283, April 20, 2023. https://www.rstreet.org/research/flexible-pro-innovation-governance-strategies-for-artificial-intelligence.

[14] Adam Thierer, *Permissionless Innovation: The Continuing Case for Comprehensive Technological Freedom*, 2nd ed (Mercatus Center at George Mason University, 2016), pp. 26-29.

[15] Gianclaudio Malgieri and Frank A. Pasquale, "From Transparency to Justification: Toward Ex Ante Accountability for AI," *Brooklyn Law School Legal Studies Paper* 712 (May 23, 2022), pp. 2-27. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4099657.

[16] Margot E. Kaminski, "Regulating the Risks of AI," *Boston University Law Review* 103 (2023), pp. 1-83. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4195066.

[17] Martin Rees, *On the Future: Prospects for Humanity* (Princeton University Press, 2018), p. 136; Thierer, *Permissionless Innovation*.

[18] Adam Thierer, "Failing Better: What We Learn by Confronting Risk and Uncertainty," in Sherzod Abdukadirov, ed., *Nudge Theory in Action: Behavioral Design in Policy and Markets* (Palgrave Macmillan, 2016), pp. 65-94.

over the last half-century, "regulation clobbered the learning curve" for many important technologies in the United States, especially nuclear, nanotechnology and advanced aviation.[19]

Although precautionary principle-based approaches would generally hamper innovation, they may be needed for specific AI applications that pose serious risks, such as pervasive automated policing efforts involving facial recognition.[20] However, as noted below, context and circumstances are essential when debating these issues, as policymakers cannot ignore the trade-offs associated with possible global regulatory schemes for AI. With this context in mind, this report broadly considers two key questions:

1. **How practical are global regulatory control efforts?**
2. **Could efforts to address global AI risks give rise to different—and even bigger—existential risks in the process?**

Many proponents of global control systems for AI or robotics often sidestep these questions, preferring to propose grand regulatory schemes enforced by enlightened global bureaucracies. But it is more likely that bureaucracy will need to give way to adhocracy. Solutions will need to be cobbled together on the fly, and continuous coordination and dialogue will be crucial. Extreme solutions—including global bans or mass surveillance regimes—must be rejected as unworkable and unwise. Finally, even for situations in which more precautionary approaches might be warranted, voluntary and nongovernmental "soft law" mechanisms could play a major role and may be the only options available.[21]

## Defining and Prioritizing Existential Risks

The definitional challenges around AI—and how to delineate AI-related risk in particular—greatly complicate governance questions. Terms like "AI risk" or "existential risk" are often casually used by those worried about the power of algorithmic or robotic systems, yet those parties frequently fail to define these terms with the level of precision needed to regulate computational systems in a meaningful way.

The term "existential" historically derives its meaning in relation to the continued existence of humanity; this study posits that the term retains the most impact when it maintains this connotation. Some technology critics use the term existential when discussing whether a social media site is undermining society.[22] Although social media and algorithms can give rise to legitimate risks, using the term too casually in this context trivializes it and can represent a form of what scholars

Terms like "AI risk" or "existential risk" are often casually used by those worried about the power of algorithmic or robotic systems, yet those parties frequently fail to define these terms with the level of precision needed to regulate computational systems in a meaningful way.

[19] J. Storrs Hall, *Where Is My Flying Car*? (Stripe Press, 2021).

[20] Matthew Feeney, "Statement for the Record, Hearing on "Facial Recognition Technology: Examining Its Use by Law Enforcement," CATO Institute, July 13, 2021. https://www.cato.org/testimony/facial-recognition-technology-examining-its-use-law-enforcement.

[21] Gary E. Marchant, "Governance of Emerging Technologies as a Wicked Problem," *Vanderbilt Law Review* 73:6 (Dec. 22, 2020), p. 1866. https://vanderbiltlawreview.org/lawreview/2020/12/governance-of-emerging-technologies-as-a-wicked-problem.

[22] Franklin Foer, *World Without Mind: The Existential Threat of Big Tech* (Penguin Press, 2017).

call "threat inflation," or "the attempt … to create concern for a threat that goes beyond the scope and urgency that a disinterested analysis would justify."[23] This could also divert attention or resources inappropriately.[24] It is essential, therefore, to appreciate the differences between various AI-related risks and understand why they do not all warrant being classified as existential.

It is essential to appreciate the differences between various AI-related risks and understand why they do not all warrant being classified as existential.

We also cannot define risk without first defining harm, but it is notoriously difficult to define harm in this context—and even more difficult should we expect different countries to agree on which harms and risks should be prioritized over others, as even experts in the field disagree. For example, some scholars and policymakers are concerned about how authoritarian regimes might try to "brainwash" people domestically or globally with AI-generated propaganda.[25] Misinformation, or so-called "deepfakes," might be one manifestation of this problem.[26] On the other hand, some free-speech advocates protest efforts to use sweeping controls to address "disinformation" because government officials would first need to define the term. This tension was evident in the United States in 2022 when a heated debate erupted over a Biden administration effort to create a Disinformation Governance Board within the Department of Homeland Security, and especially the question of how they would define online disinformation.[27] These types of definitional disputes would be even more controversial and complex at the global level.

These concerns can sometimes be overblown, leading to irrational catastrophizing that can pose a risk in and of itself.[28] Indeed, the greatest of all risks lies in efforts to avoid risk altogether. A policy preference for the status quo (i.e., the precautionary principle) is a recipe for technological and economic stasis. As noted in one study, "living in constant fear of worst-case scenarios—and premising public policy on them—means that best-case scenarios will never come about."[29] The optimal solution to technological risk often lies in more technological innovation to overcome those problems, not less. It is irresponsible to suggest that we should stop all algorithmic or robotic innovation in the name of limiting AI-related risks.[30] "Cultures that attempt to block technology for reasons that appear desirable will … eventually be dominated by those that embrace it," argues a scientist at Arizona State University.[31]

[23] Jane K. Cramer and A. Trevor Thrall, "Introduction: Understanding Threat Inflation," in A. Trevor Thrall and Jane K. Cramer, eds., *American Foreign Policy and the Politics of Fear: Threat Inflation Since 9/11* (Routledge, 2009), p. 1; Adam Thierer, "Technopanics, Threat Inflation, and the Danger of an Information Technology Precautionary Principle," *Minnesota Journal of Law, Science & Technology* 14:1 (2013), pp. 312-350. https://scholarship.law.umn.edu/mjlst/vol14/iss1/8.

[24] "Existential Risk: Diplomacy and Governance," Global Priorities Project, Feb. 3, 2017, p. 6. http://globalprioritiesproject.org/2017/02/existential-risk-diplomacy-and-governance.

[25] Bill Drexel and Caleb Withers, "Generative AI could be an authoritarian breakthrough in brainwashing," *The Hill*, Feb. 26, 2023. https://thehill.com/opinion/technology/3871841-generative-ai-could-be-an-authoritarian-breakthrough-in-brainwashing.

[26] Daniel L. Byman et al., "Deepfakes and international conflict," Brookings, January 2023. https://www.brookings.edu/research/deepfakes-and-international-conflict.

[27] Adam Thierer and Patricia Patnode, "Disinformation About the Real Source of the Problem," RealClearPolicy, May 23, 2022. https://www.realclearpolicy.com/articles/2022/05/23/disinformation_about_the_real_source_of_the_problem_833681.html.

[28] James R. Ostrowski, "Shallowfakes," *The New Atlantis* (Spring 2023). https://www.thenewatlantis.com/publications/shallowfakes.

[29] Thierer, *Permissionless Innovation*, p. 2.

[30] Scott Alexander, "Why Not Slow AI Progress?," Astral Codex Ten, Aug. 8, 2022. https://astralcodexten.substack.com/p/why-not-slow-ai-progress.

[31] Braden Allenby, "The Dynamics of Emerging Technology Systems," in Gary E. Marchant et al., eds., *Innovative Governance Models for Emerging Technologies* (Edward Elgar, 2013), p. 33.

Policymakers and scholars should also be extremely cautious about the language they use to describe new classes of technologies, lest they cast too wide of a net with calls for controlling weapons of mass destruction that may be nothing of the sort. Suggestions that every new technology poses a catastrophic or existential risk will desensitize people to legitimate risks that may be associated with a narrower class of innovations. Thus, instead of using fear-based appeals and advocating for extreme (and likely unworkable) global regulatory schemes, it would be wiser to build on existing laws, norms and alternative governance frameworks.

## Challenges Associated with Global Regulatory Regimes

AI has begun to raise a variety of national security issues, and scholars and policymakers have started thinking about how "AI arms control" might work in practice.[32] Some have called for global controls for AI through a global regulatory authority, such as a hypothetical International AI Organization.[33] But just as chemical and nuclear preventive arms-control efforts have long been haunted by enforcement challenges, imposing limits on dangerous forms of AI or robotics would also be complicated and contentious.[34]

Nick Bostrom, Director of the Future of Humanity Institute at the University of Oxford, has done the most important work on AI existential risk and its potential global regulation. He has written extensively about the dangers of superintelligence and what he calls the "vulnerable world hypothesis."[35] While many proposals for global AI control tend to be highly aspirational and lack specific details, Bostrom has outlined a variety of specific regulatory options for addressing these concerns. In addition, many current discussions about global AI control point back to Bostrom's proposals, which suggests that they could form the groundwork for future regulatory approaches.

Bostrom argues that "[o]ur approach to existential risks cannot be one of trial-and-error," explaining that some theoretical risks are so potentially catastrophic that permissionless innovation is no longer the optimal policy default. But that does not automatically mean that the precautionary principle should be the default instead.

Bostrom argues that "[o]ur approach to existential risks cannot be one of trial-and-error," explaining that some theoretical risks are so potentially catastrophic that permissionless innovation is no longer the optimal policy default.[36] But that does not automatically mean that the precautionary principle should be the default instead. As Bostrom notes, "stopping technological development would require something close to a cessation of inventive activity everywhere in the world," which is not realistic and would constitute its own existential catastrophe.[37] Still, that position has not stopped other scholars from openly questioning whether we should temporarily pause AI

[32] Matthew Mittelsteadt, "AI Verification: Mechanisms to Ensure AI Arms Control Compliance," Center for Security and Emerging Technology, February 2021. https://securitypolicylaw.syr.edu/wp-content/uploads/2021/02/Mittelstaedt_AI_Verification_2021.pdf.

[33] Erdélyi and Goldsmith, pp. 95-101. https://dl.acm.org/doi/10.1145/3278721.3278731.

[34] Myriam Dunn Cavelty et al., "'Killer Robots' and Preventive Arms Control," in *The Routledge Handbook of Security Studies*, 2nd Edition (Routledge Hardback, 2016), pp. 457-468.

[35] Nick Bostrom, "The Vulnerable World Hypothesis," *Global Policy* 10:4 (November 2019), pp. 455-476. https://nickbostrom.com/papers/vulnerable.pdf.

[36] Nick Bostrom, "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazard," *Journal of Evolution and Technology* 9:1 (2002). https://nickbostrom.com/existential/risks.html.

[37] Bostrom, "The Vulnerable World Hypothesis," p. 462. https://nickbostrom.com/papers/vulnerable.pdf.

development, and some media pundits have argued that, "[p]umping the brakes on artificial intelligence could be the best thing we ever do for humanity."[38]

Bostrom takes a more measured approach, suggesting "*limited* curtailments of inventive activities."[39] He proposes the "Principle of Differential Technological Development," which would "[r]etard the development of dangerous and harmful technologies, especially ones that raise the level of existential risk; and accelerate the development of beneficial technologies, especially those that reduce the existential risks posed by nature or by other technologies."[40] While he admits that it is strategically difficult to effectively implement differential technological development, he nonetheless believes it should be attempted if for no other reason than to buy some additional time, which is the same logic that inspired the Future of Life Institute's call for a pause on large-scale AI experiments.[41]

There is some precedent for this approach. After World War I and World War II, governments and other groups came together to address the dangerous uses of chemical and nuclear technologies. Shortly after World War I, the "Geneva Protocol for the Prohibition of the Use in War of Asphyxiating, Poisonous or other Gases, and of Bacteriological Methods of Warfare" was formulated to limit the uses of chemical weapons in future conflicts.[42] Similarly, after World War II, international treaties and other agreements limited the possession or enrichment of uranium, as well as the ability to build or traffic nuclear weapons. In addition, the "Treaty on the Non-Proliferation of Nuclear Weapons" (NPT) was created in 1968 to advance the peaceful uses of nuclear technology and limit its dangerous applications and is supported by the International Atomic Energy Agency's (IAEA) mission to "accelerate and enlarge the contribution of atomic energy to peace, health and prosperity throughout the world."[43]

Bostrom takes a more measured approach, suggesting "*limited* curtailments of inventive activities." He proposes the "Principle of Differential Technological Development," which would "[r]etard the development of dangerous and harmful technologies, especially ones that raise the level of existential risk."

Fortunately, humanity's worst fears about chemical and nuclear weapons have so far been avoided; although more than 63,000 nuclear weapons existed in the 1980s, that number has dropped to less than 14,000 today.[44] While this is still a dangerous number of nuclear weapons, the diminished number demonstrates the effective international effort to mitigate their creation and use. Yet it is impossible to quantify how much of this success should be attributed to these global efforts versus the extremely costly process of obtaining or producing such weapons. This is a key question, as material costs would not be a constraining factor in the development and spread of computational technologies and algorithmic applications that lack a physical form.

[38] Michelle Rempel Garner and Gary Marcus, "Is it time to hit the pause button on AI?," Michelle Rempel Garner, Feb. 26, 2023. https://michellerempelgarner.substack.com/p/is-it-time-to-hit-the-pause-button; Sigal Samuel, "The case for slowing down AI," *Vox*, March 20, 2023. https://www.vox.com/the-highlight/23621198/artificial-intelligence-chatgpt-openai-existential-risk-china-ai-safety-technology.

[39] Bostrom, "The Vulnerable World Hypothesis," p. 462. https://nickbostrom.com/papers/vulnerable.pdf.

[40] Ibid.

[41] Ibid., p. 463; "Pause Giant AI Experiments: An Open Letter," Future of Life Institute, March 22, 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments.

[42] "Protocol for the Prohibition of the Use in War of Asphyxiating, Poisonous or Other Gases, and of Bacteriological Methods of Warfare," United Nations Office for Disarmament Affairs, June 17, 1925. https://www.un.org/disarmament/wmd/bio/1925-geneva-protocol.

[43] "Statute: As Amended up to 28 December 1989," International Atomic Energy Agency, p. 5. https://www.iaea.org/sites/default/files/statute.pdf.

[44] Peter Wildeford, "Human survival is a policy choice," Pasteur's Cube, June 2, 2022. https://www.pasteurscube.com/surviving-into-the-future-is-a-policy-choice.

Nonetheless, Bostrom proposes five specific control mechanisms for such technologies:

1. **Prevent dangerous information from spreading**
2. **Restrict access to requisite materials, instruments and infrastructure**
3. **Deter potential evildoers by increasing the chance of their getting caught**
4. **Be more cautious and do more risk-assessment work**
5. **Establish some kind of surveillance and enforcement mechanism that would make it possible to interdict attempts to carry out destructive acts**

FAKE NEWS
TECHNOLOGY CONTROL MECHANISMS
****
RISK

Of course, these measures all have serious trade-offs and limitations of their own, and some of these solutions may work better for particular algorithmic or robotic risks than others.

## Controlling the Use of "Killer Robots"

To move away from the theoretical and toward the practical, we can look at the concerns regarding the development of robotic military technologies or LAWS in the context of Bostrom's framework.

One of the most immediately concerning classes of LAWS is so-called "killer robots," which could include not only autonomous robot soldiers but also other types of LAWS, such as drones, drone swarms, or even technology like mechanical dogs equipped with weapons that could act as "slaughterbots" or autonomous robotic weapon systems programmed to conduct assassinations or engage in other types of terrorist activity.[45]

Because these dystopian scenarios have been explored in countless science fiction books, films and television shows, they tend to drive much of the public imagination—and policy discussions—surrounding AI.[46] Such depictions make it easy to wonder how we might prevent *Terminator*-esque scenarios to the extent that such risks are theoretically possible. These are technological risks that everyone would agree are tangible, irreversible and catastrophic in nature, and some degree of precautionary principle-based regulation may be required to avoid them.

LAWS are technological risks that everyone would agree are tangible, irreversible and catastrophic in nature, and some degree of precautionary principle-based regulation may be required to avoid them

However, the challenge in applying even a reasonable degree of precautionary principle-based regulation as outlined in Bostrom's framework is that, much like chemical and nuclear technologies before them, robots and AI can be dual-use technologies that have many potential beneficial uses.[47] Much of the progress in this field is driven by civilian-based innovations (e.g., self-driving cars, household robots).[48] Thus, the same computational engines that give the world universal language

[45] Anzhelika Solovyeva and Nik Hynek, "Going Beyond the «Killer Robots» Debate: Six Dilemmas Autonomous Weapon Systems Raise," *Central European Journal of International and Security Studies* 12:3 (2018), pp. 166-208. https://www.cejiss.org/going-beyond-the-killer-robots-debate-six-dilemmas-autonomous-weapon-systems-raise-0; Zachary Kallenborn and Philipp C. Bleek, "Swarming destruction: drone swarms and chemical, biological, radiological, and nuclear weapons," *The Nonproliferation Review* 25:5-6 (2018), pp. 523-543. https://www.tandfonline.com/doi/abs/10.1080/10736700.2018.1546902; Stuart Russell et al., "Why You Should Fear 'Slaughterbots'—A Response," *IEEE Spectrum*, Jan. 23, 2018. https://spectrum.ieee.org/why-you-should-fear-slaughterbots-a-response.

[46] Adam Thierer, "How Science Fiction Dystopianism Shapes the Debate over AI & Robotics," *Discourse*, July 26, 2022. https://www.discoursemagazine.com/culture-and-society/2022/07/26/how-science-fiction-dystopianism-shapes-the-debate-over-ai-robotics; Anne Hobson, "Westworld shouldn't frame debate over artificial intelligence," *The Hill*, March 8, 2017. https://www.rstreet.org/2017/03/08/westworld-shouldnt-frame-debate-over-artificial-intelligence.

[47] Brundage et al., p. 16. https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf.

[48] Cavelty et al., p. 465.

translation—a clear boon to those seeking to communicate across the globe—could also give us the ability to create misinformation or deepfakes, which could be greatly destabilizing to global trust and public order.[49] In short, if used improperly, some of these systems could have disastrous consequences that are entirely disconnected from the innovative goals of their creation.

Still, the threat of killer robots—especially in the context of drones, which have already been used extensively in modern warfare—provides us with a more concrete target for discussion and possible policy action.[50] Two key AI existential risks that we can assess as part of this discussion are: **(1) whether the concept of "compute" can be meaningfully controlled** and **(2) the challenge of mass surveillance solutions**.

Assessment of Key AI Existential Risk #1

## The Challenge of Controlling "Compute"

As with firearms and most other weapons technologies before them, drones and other robotic LAWS are technological artifacts that are both tangible and trackable.[51] By comparison, calls for global governance steps to address AGI or machine-based superintelligence represent a far more open-ended and ambitious challenge.[52] As the recent Future of Life Institute open letter makes clear, stopping or slowing AGI ultimately comes down to regulating "compute," which refers to the enormous computing capabilities that drive powerful algorithmic systems.[53] Although many physical computing systems are needed here, too, digital code—which is intangible information—lies at the heart of this technological process. The quicksilver nature of digital code significantly complicates regulatory schemes aimed at controlling the development of more powerful computations systems.

The quicksilver nature of digital code significantly complicates regulatory schemes aimed at controlling the development of more powerful computations systems.

But this fact also complicates efforts to control physical robotic systems to some extent because today's AI research can be undertaken in a domestic setting with standard consumer hardware.[54] As a result, placing a ban on killer robots would do little to slow or stop the AI arms race.[55] As a Brookings scholar explains:

> AI is very different from a domain like nuclear weapons development, where compliance with moratoriums is feasible (though not always easy) to track because the associated materials and technologies, such as uranium and nuclear centrifuges, are hard to come by, difficult to work with, and have a very limited set of uses. With AI systems, the key ingredients are data and computing power, both of which are readily accessible and have an essentially limitless list of non-moratorium-violating uses.[56]

[49] Byman et al. https://www.brookings.edu/research/deepfakes-and-international-conflict.

[50] Eric Schmidt and Jared Cohen, *The New Digital Age: Reshaping the Future of People, Nations and Business* (Alfred A. Knopf, 2013), p. 201; Peter Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century* (Penguin, 2009).

[51] Martin Van Creveld, *Technology and War: From 2000 B.C. to the Present* (The Free Press, 1989).

[52] Joseph Carlsmith, "Is Power-Seeking AI an Existential Risk?," *Computers and Society* (April 2021). https://arxiv.org/abs/2206.13353.

[53] Andrew Lohn and Micah Musser, "AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?," Center for Security and Emerging Technology (January 2022). https://cset.georgetown.edu/publication/ai-and-compute.

[54] Edward Moore Geist, "It's already too late to stop the AI arms race—We must manage it instead," *Bulletin of the Atomic Scientists* 72:5 (2016), p. 320. https://www.tandfonline.com/doi/full/10.1080/00963402.2016.1216672.

[55] Ibid., p. 319.

[56] John Villasenor, "The problems with a moratorium on training large AI systems," Brookings, April 11, 2023. https://www.brookings.edu/blog/techtank/2023/04/11/the-problems-with-a-moratorium-on-training-large-ai-systems.

More importantly, the Future of Life open letter to pause powerful AI systems comes at a time when many governments are calling for significant expansions in computing power for AI and quantum science.[57] In the United States, the only major technology legislation Congress enacted recently was the Chips and Science Act, which is aimed at advancing computational capabilities through significant federal investments "to keep the United States the leader in the industries of tomorrow, including nanotechnology, clean energy, quantum computing, and artificial intelligence."[58] The Chips Act followed earlier efforts by the Obama and Trump administrations to promote AI and quantum science through various initiatives. A Stanford University researcher notes that the United States continues to invest in AI research and development, "In fiscal year 2022, U.S. government agencies allocated $1.7 billion to AI R&D, up 13% from the year prior and an increase of 209% from 2018 [and] the Department of Defense, in its nonclassified AI budget request, asked for $1.1 billion, a 26% increase from the prior year."[59]

The Future of Life open letter to pause powerful AI systems comes at a time when many governments are calling for significant expansions in computing power for AI and quantum science.

"In fiscal year 2022, U.S. government agencies allocated **$1.7 billion** to AI R&D, **up 13%** from the year prior."

Many other global governments are looking to do the same and are also making significant investments in their technological base—both in terms of broad-based computational capabilities and target algorithmic technologies and sectors.[60] In March 2023, on the same day that the U.K.-based Future of Life was issuing its call for a global AI pause, the U.K. government issued a comprehensive new AI policy vision "to turbocharge growth."[61] When announcing the plan, the U.K. Secretary of State for Science, Innovation and Technology noted that the country had invested over £2.5 billion in AI since 2014 and had recently announced allocating £110 million for an AI Tech Missions Fund and £900 million to establish a new "AI Research Resource and to develop an exascale supercomputer capable of running large AI models – backed up by our new £8 million AI Global Talent Network and £117 million of existing funding to create hundreds of new PhDs for AI researchers."[62]

Meanwhile, China continues to make massive investments in its AI and robotic capabilities, and many other countries are making significant investments in these fields as well.[63] Importantly, reported expenditures of these countries

[57] Gil Press, "New Funding For Quantum Computing Accelerates Worldwide," *Forbes*, Jan. 31, 2023. https://www.forbes.com/sites/gilpress/2023/01/31/new-funding-for-quantum-computing-accelerates-worldwide/?sh=19f76c04b35b.

[58] "FACT SHEET: CHIPS and Science Act Will Lower Costs, Create Jobs, Strengthen Supply Chains, and Counter China," The White House, Aug. 9, 2022. https://www.whitehouse.gov/briefing-room/statements-releases/2022/08/09/fact-sheet-chips-and-science-act-will-lower-costs-create-jobs-strengthen-supply-chains-and-counter-china.

[59] Shana Lynch, "2023 State of AI in 14 Charts," Stanford University Human-Centered Artificial Intelligence, April 3, 2023. https://hai.stanford.edu/news/2023-state-ai-14-charts.

[60] "The Global AI Index," Tortoise, last accessed April 7, 2023. https://www.tortoisemedia.com/intelligence/global-ai.

[61] "UK unveils world leading approach to innovation in first artificial intelligence white paper to turbocharge growth," Gov.UK, March 29, 2023. https://www.gov.uk/government/news/uk-unveils-world-leading-approach-to-innovation-in-first-artificial-intelligence-white-paper-to-turbocharge-growth.

[62] "A pro-innovation approach to AI regulation," Gov.UK, March 29, 2023. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

[63] Kai Shen et al., "The next frontier for AI in China could add $600 billion to its economy," QuantumBlack AI by McKinsey, June 7, 2022. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-next-frontier-for-ai-in-china-could-add-600-billion-to-its-economy; Yi Wu, "AI in China: Regulations, Market Opportunities, Challenges for Investors," China Briefing, Oct. 14, 2022. https://www.china-briefing.com/news/ai-in-china-regulatory-updates-investment-opportunities-and-challenges; Akira Oikawa and Niki Mizuguchi, "Japan quantum computer debut sets off scramble for tech breakthroughs," NikkeiAsia, March 28, 2023. https://asia.nikkei.com/Business/Technology/Japan-quantum-computer-debut-sets-off-scramble-for-tech-breakthroughs.

are merely the official public statements the governments have made about their investments. It is impossible to know exactly how much governments are spending on covert projects or capabilities—a fact that is equally true of the U.S. government. Regardless, even just the known investment activity of various nations makes it clear that the calls to have governments restrict their aggregate computational capabilities will be ignored.

## Could AI-Enabled Systems Save Lives in Armed Conflicts?

**It is worth noting that some scholars have made the case that robotic or autonomous systems could help diminish human casualties and suffering during armed conflict.[1] AI-enabled military systems might also make better judgments (such as more precision strikes) during conflict than humans, who may make emotional decisions that cost more lives than necessary. In addition, the increased use of autonomous systems could help reduce dependence on large-scale and more expensive weapon systems, including dangerous nuclear stockpiles, other traditional munitions (like land mines) and other indiscriminate methods of warfare (like carpet bombing). The reduced use of the last two, in particular, could help minimize civilian and other casualties. Some scholars argue that these factors create a powerful moral obligation to use more, not less, robotic technology in armed conflict or to address other military needs.[2]**

**On the other hand, even if some lives are spared in the short-term, robotic armies could make warfare less expensive and easier to pursue in an even more remote fashion—including using robots as distant stand-ins for humans, thus potentially giving rise to greater long-term risk for civilization though "killing made easy."[3] There is no way to determine how this cost-benefit test might play out in advance, but those concerned about this possibility point out that it is worth considering deterrence efforts to avoid widespread robotic conflict or "runaway AI" scenarios involving LAWS.**

[1] Steven Umbrello et al., "The future of war: could lethal autonomous weapons make conflict more ethical?," AI & Society 35 (Feb. 6, 2019), pp. 273-282. https://link.springer.com/article/10.1007/s00146-019-00879-x.

[2] Ronald C. Arkin, "The Case for Ethical Autonomy in Unmanned Systems," Journal of Military Ethics 9:4 (2010), pp. 332-341. https://www.tandfonline.com/doi/abs/10.1080/15027570.2010.536402.

[3] Noel Sharkey, "Killing Made Easy: From Joysticks to Politics," in Patrick Lin et al., eds., Robot Ethics: The Ethical and Social Implications of Robotics (The MIT Press, 2012), p. 111.

## The Problems with Mass Surveillance Solutions

Assessment of Key AI Existential Risk #2

The second concrete target for discussion and possible policy action is the problem with mass surveillance. Returning to the final of Bostrom's five proposals for addressing existential risk—preventive policing through stronger intra- and interstate governance—it is clear that significant new issues arise with stronger preventive policing policies such as surveillance. As one journalist observed, "if the continued survival of humanity depended on successfully imposing worldwide surveillance," it could "lead to disastrous unintended consequences."[64] Bostrom

[64] Kelsey Piper, "How technological progress is making it likelier than ever that humans will destroy ourselves," *Vox*, Nov. 19, 2018. https://www.vox.com/future-perfect/2018/11/19/18097663/nick-bostrom-vulnerable-world-global-catastrophic-risks.

himself admits that there are massive risks and downsides to global governance and mass surveillance but suggests it is worth the sacrifice to save humanity.[65] Others have suggested similar steps might be needed, including a computing surveillance and tracking regime complete with backdoors for government data access.[66]

A mass surveillance and computer tracking apparatus would not necessarily guarantee workable containment solutions for the sort of catastrophes critics fear, but it would open the door to a different type of disaster in the form of highly repressive state controls on communications, individual movement and other activities. As one analyst comments, "[g]lobal totalitarianism is its own existential risk."[67] Of note, Bostrom's proposal is reminiscent of a proposal floated by U.S. defense officials in the late 1990s to create government-mandated backdoors for encrypted digital systems and applications to allow government access in the name of advancing national security and law enforcement goals.[68]

Topping the list of mass surveillance issues is the question of who would set up, fund and administer such a global regime. Practically speaking, it would be impossible to ensure that the surveillance infrastructure covered every part of the world where risky research and development efforts might be undertaken.

Ignoring the specter of mass surveillance associated with such a proposal, other costs and resource constraints remain a serious problem for the sort of global regulatory regime Bostrom and others recommend. Topping the list of issues is the question of who would set up, fund and administer such a global surveillance regime. There is little reason to believe that the United Nations (U.N.) or any other global body could convince the leaders of every nation to let them create a worldwide panopticon that has access, in real time, to all scientific developments happening in their countries. Practically speaking, it would be impossible to ensure that the surveillance infrastructure covered every part of the world where risky research and development efforts might be undertaken. Some scholars have criticized Bostrom for not adequately weighing these various downsides in his framework.[69]

Finally, mass surveillance schemes could discourage research into a great many risk-reducing technological applications and thus undermine Bostrom's other goal of "accelerat[ing] the development of beneficial technologies, especially those that reduce the existential risks posed by nature or by other technologies."[70] Scientists cannot be monitored perpetually to stop them from developing just bad robotic capabilities; they would need to be monitored in real time to evaluate all the potential uses of what they are working on. Moreover, rogue actors (whether they be state or nonstate actors) would not likely abide by such restrictions—even if they make pledges to do so.

[65] Kristin Houser, "Professor: Total Surveillance Is the Only Way to Save Humanity," Futurism, April 19, 2019. https://futurism.com/simulation-mass-surveillance-save-humanity.

[66] Luke Muehlhauser, "12 tentative ideas for US AI policy," Open Philanthropy, April 17, 2023. https://www.openphilanthropy.org/research/12-tentative-ideas-for-us-ai-policy.

[67] Maxwell Tabarrok, "Enlightenment Values in a Vulnerable World," Effective Altruism Forum, July 18, 2022. https://forum.effectivealtruism.org/posts/A4fMkKhBxio83NtBL/enlightenment-values-in-a-vulnerable-world.

[68] Sean Gallagher, "What the government should've learned about backdoors from the Clipper Chip," *Ars Technica*, Dec. 14, 2015. https://arstechnica.com/information-technology/2015/12/what-the-government-shouldve-learned-about-backdoors-from-the-clipper-chip.

[69] Robin Hanson, "Vulnerable World Hypothesis," Overcoming Bias, Nov. 16, 2018. http://www.overcomingbias.com/2018/11/vulnerable-world-hypothesis.html.

[70] Bostrom, "The Vulnerable World Hypothesis," p. 462. https://nickbostrom.com/papers/vulnerable.pdf.

This is all equally true for the underlying computing power behind large-scale algorithmic systems and specific robotic applications. The basic science behind good and bad uses of robotics is largely the same. As a result, the beneficial applications they might develop could be discouraged if they feared a loss of their intellectual property, trade secrets, or negative blowback from bureaucrats monitoring and evaluating their activities from afar. The chilling effect would be very real, and the economic, scientific and security consequences would be deleterious. For these reasons, calls for global surveillance regimes to police AI risks are both highly unworkable and dangerous.

The beneficial applications researchers might develop could be discouraged if they feared a loss of their intellectual property, trade secrets, or negative blowback from bureaucrats monitoring and evaluating their activities. The economic, scientific and security consequences would be deleterious.

Bostrom's three other options might have greater merit as applied to killer robots and their control. To recap, those other options include restrictions on access to key materials or dangerous building blocks central to these applications; deterrence efforts aimed at finding potential evildoers, increasing the likelihood of them getting caught; and more general efforts by various actors to exercise more caution and do more and better risk assessment work. Overall, however, the exploration of Bostrom's suggestions enables the consideration of specific AI development pathways and the regulatory complications and needs that they could engender. As such, international officials might be able to utilize some of these ideas and proposals to address dangerous AI developments.

## The Limitations of International Treaties and Accords

Many international efforts are already underway to address lethal or military uses of robots. The International Committee for Robot Arms Control is a nongovernmental organization (NGO) formed in 2009 to push for global regulation of these technologies. Likewise, the Campaign to Stop Killer Robots, which launched in 2013, seeks a multinational treaty to ban fully autonomous weapons.[71] Moreover, before it released its recent open letter calling for a 6-month pause on powerful AI systems, the Future of Life Institute had released another open letter in 2018 called the "Lethal Autonomous Weapons Pledge," which was signed by hundreds of organizations and thousands of individual experts.[72] The letter calls upon governments and government leaders to stand together against LAWS with international laws, regulations, and norms, and signatories vow that they will not support or participate in their development, manufacture, use or trade.[73] Both open letters from The Future of Life Institute flowed out of earlier declarations issued in its "Asilomar AI Principles"—a set of proclamations and recommendations drafted by experts at the 2017 Asilomar Beneficial AI conference.[74]

[71] "About Us," Stop Killer Robots, last accessed April 27, 2022. https://www.stopkillerrobots.org/about.

[72] "Lethal Autonomous Weapons Pledge," Future of Life Institute, June 6, 2018. https://futureoflife.org/lethal-autonomous-weapons-pledge.

[73] Ibid.

[74] "AI Principles," The Future of Life Institute, Aug. 11, 2017. https://futureoflife.org/open-letter/ai-principles.

It remains unclear whether these calls for action or pledges will influence the development of more formal international policies. It is equally unclear how treaty enforcement would work or whether the documents would be binding in any meaningful sense. In theory, at least, the sale or trade of killer robot applications could be somewhat limited through international accords and actions, perhaps using the Geneva Protocol for chemical weapons or the NPT and the IAEA for nuclear proliferation as models.[75] The U.N.'s 1972 Biological Weapons Convention (BWC), which sought to ban biological and toxic weapons globally, and its 1981 Convention on Prohibitions or Restrictions on the Use of Certain Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, which covers mines, booby traps, lasers, and other weapons, may provide another deterrence model, though they lack any real enforcement mechanisms.[76] Finally, the International Criminal Court, whose mission is to hold responsible parties accountable for crimes and to prevent such crimes in the future, could also play a role in addressing lethal uses of emerging technologies.[77]

Of course, over time, the body of laws, accords and general principles that make up the law of armed conflicts will evolve to accommodate these new technological capabilities, but these systems have many practical shortcomings, which are discussed below.

## Noncompliance with Nonproliferation Requirements

A primary focus of the previous international treaties and accords was nonproliferation, or efforts to limit the development or spread of certain weapons systems or applications that hold the potential to do catastrophic harm.[78] But according to one scholar, "[n]onproliferation was always more dream than practical program."[79] Obviously, nonstate actors (such as terrorist groups) will not agree to be bound by such restrictions.[80] There is also the practical problem of rogue states or nations that refuse to abide by the terms of such agreements and treaties even after signing them. This is the issue the world faces with nuclear nonproliferation efforts—and not just with states like North Korea and Iran. For example, although it signed the 1972 BWC, the former Soviet Union immediately ignored the treaty and began to secretly develop biological weapons on a massive scale.[81] South Africa and Iraq were also later revealed to have ignored the treaty.[82]

According to one scholar, "[n]onproliferation was always more dream than practical program." There is also the practical problem of rogue states or nations that refuse to abide by the terms of such agreements and treaties even after signing them.

[75] Mauricio Baker, "Nuclear Arms Control Verification and Lessons for AI Treaties," arXiv, April 8, 2023. https://arxiv.org/abs/2304.04123.

[76] Robert J. Mathews, "The 1980 Convention on Certain Conventional Weapons: A useful framework despite earlier disappointments," *International Review of the Red Cross* 83:844 (December 2001), pp. 991-1012. https://www.cambridge.org/core/journals/international-review-of-the-red-cross/article/abs/1980-convention-on-certain-conventional-weapons-a-useful-framework-despite-earlier-disappointments/70C1E6096B9E145FB27875820DC66EA3.

[77] "About the Court," International Criminal Court, last accessed April 12, 2023. https://www.icc-cpi.int/about.

[78] Justin Anderson et al., "Nonproliferation and Counterproliferation," Oxford Bibliographies, March 27, 2014. https://www.oxfordbibliographies.com/display/document/obo-9780199743292/obo-9780199743292-0026.xml.

[79] Walter Russell Mead, "How Obama Killed Nuclear Nonproliferation," *The Wall Street Journal*, April 10, 2023. https://www.wsj.com/articles/how-obama-killed-nuclear-nonproliferation-npt-soviet-union-ukraine-deterrence-bill-clinton-russia-invasion-rules-based-order-49959cc8.

[80] Braden R. Allenby, *The Rightful Place of Science: Future Conflict & Emerging Technologies* (Consortium for Science, Policy, & Outcomes, 2016), pp. 15-18.

[81] Geist, p. 320. https://www.tandfonline.com/doi/full/10.1080/00963402.2016.1216672.

[82] Toby Ord, *The Precipice: Existential Risk and the Future of Humanity* (Hachette Books, 2020), p. 136.

The fundamental distrust that exists between many major geopolitical powers is unlikely to change in the era of AI superpowers.[83] Those calling for outright bans on AI or LAWS often adopt the extreme position that any and all development of new technologies should stop.[84] Although the cost-benefit ratio of that position could be debated, a George Mason University economist argues that "China, Russia, and many other rival nations have no such plans, and the U.S. has no real choice other than to try to stay ahead of them."[85]

Russia has already stated that it has no intention of complying with a ban on killer robots.[86] The United States has also rejected calls for outright bans on killer robots or LAWS, preferring instead to push for some sort of code of conduct to govern their use.[87] Australia, Israel and the U.K. are also opposed to an outright ban.[88] In many ways, the rejection of a call for a global ban on killer robots is mirroring what happened with the Mine Ban Treaty (or "Ottawa Treaty")—a global effort undertaken in 1997 to eradicate the production and use of land mines. Despite an ongoing global campaign and widespread media visibility, over 30 nations refused to sign the treaty, including the United States, China, Russia, India, and both North and South Korea.

It is clear that many nations will refuse to tie their hands preemptively when asked to not develop offensive AI-oriented military capabilities. Doing so could give rise to a different existential risk for these nations because, by freezing their capabilities, hostile states and rouge actors might come to possess capabilities that place them at risk.

This makes it clear that many nations will refuse to tie their hands preemptively when asked to not develop offensive AI-oriented military capabilities. Doing so could give rise to a different existential risk for these nations because, by freezing their capabilities, hostile states and rouge actors might come to possess capabilities that place them at risk. This would qualify as a type of risk substitution—the problem that develops "when one type of adverse outcome is replaced by another adverse outcome in the same target population."[89] As a result, a 2021 report by the National Security Commission on Artificial Intelligence concluded that:

> Defending against AI-capable adversaries operating at machine speeds without employing AI is an invitation to disaster. Human operators will not be able to keep up with or defend against AI-enabled cyber or disinformation attacks, drone swarms, or missile attacks without the assistance of AI-enabled machines. National security professionals must have access to the world's best technology to protect themselves, perform their missions, and defend us.[90]

[83] Kai-Fu Lee, *AI Superpowers: China, Silicon Valley, and the New World Order* (Houghton Mifflin Harcourt, 2018).

[84] Tyler Cowen, "What If Our Technology Turns Against Us?," *Bloomberg*, Feb. 6, 2022. https://www.bloomberg.com/opinion/articles/2022-02-06/what-if-our-technology-turns-against-us.

[85] Ibid.

[86] Dan Robitzki, "Russia: Our Killer Robots Don't Need Any Pesky International Laws," The Byte, Aug. 4, 2021. https://futurism.com/the-byte/russia-killer-robots-international-laws.

[87] "US rejects calls for regulating or banning 'killer robots,'" *The Guardian*, Dec. 2, 2021. https://www.theguardian.com/us-news/2021/dec/02/us-rejects-calls-regulating-banning-killer-robots.

[88] Damien Gayle, "UK, US and Russia among those opposing killer robot ban," *The Guardian*, March 29, 2019. https://www.theguardian.com/science/2019/mar/29/uk-us-russia-opposing-killer-robot-ban-un-ai.

[89] John D. Graham and Jonathan Baert Wiener, "Confronting Risk Tradeoffs," in John D. Graham and Jonathan Baert Wiener, eds., *Risk vs. Risk: Tradeoffs in Protecting Health and the Environment* (Harvard University Press, 1995), p. 25.

[90] National Security Commission on Artificial Intelligence, *Final Report*, p. 9. https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

This is not to say that multinational treaties and accords do not play an important role in helping shape global norms, even when they lack teeth. In February 2023, the U.S. Department of State issued a "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," outlining 12 best practices to guide the development and use of military AI capabilities.[91] The United States released these principles before attending the first international summit on the responsible development, deployment and use of AI in the military domain (REAIM) at The Hague that same week.[92] At the REAIM conference, over 60 countries, including the United States and China, endorsed a call to action that included a set of generic best practices for military use of AI.[93] To win widespread support, however, the document steered clear of any sweeping calls to substantive action and instead engaged in a great deal of back-and-forth about how signatories would address these concerns. For example, on one hand, the statement identified, "the importance of ensuring appropriate safeguards and human oversight of the use of AI systems, bearing in mind human limitations due to constraints in time and capacities."[94] But the document also acknowledged that, "failure to adopt AI in a timely manner may result in a military disadvantage, while premature adoption without sufficient research, testing and assurance may result in inadvertent harm."[95] This leaves a great deal of discretion to each signatory to interpret the call to action to fit their own national security or economic development needs.

Importantly, the document stressed the essential role of ongoing dialogue in addressing these concerns, outlining, "[w]e are committed to continuing the global dialogue on responsible AI in the military domain in a multi-stakeholder and inclusive manner and call on all stakeholders to take their responsibility in contributing to international security and stability in accordance with international law."[96] This sort of continuous dialogue about AI risks is perhaps the most important thing we can do to address algorithmic or robotic risks, and we will explore this strategy in more detail later in this paper.

Continuous dialogue about AI risks is perhaps the most important thing we can do to address algorithmic or robotic risks.

## The Dangers of Wishful Thinking, Abstract Values and Ambiguous Proposals

Nonetheless, it is important to not place too much stock in ambitious proposals aimed at addressing AI risks through formally binding international regulatory agreements, especially when they involve sweeping calls for pauses or bans on algorithmic or computational technologies. We must set realistic goals when considering how to approach AI risks—both domestically and globally—and avoid

[91] Bureau of Arms Control, Verification and Compliance, "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," U.S. Department of State, Feb. 16, 2023. https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy.

[92] Toby Sterling, "U.S., China, other nations urge 'responsible' use of military AI," *Reuters*, Feb. 16, 2023. https://www.reuters.com/business/aerospace-defense/us-china-other-nations-urge-responsible-use-military-ai-2023-02-16.

[93] "REAIM Call to Action," Responsible AI in the Military Domain Summit, Feb. 15-16, 2023. https://www.mofa.go.jp/mofaj/files/100465549.pdf.

[94] Ibid.

[95] Ibid.

[96] Ibid.

head-in-the-sand "safetyism" beliefs that good intentions, aspirational principles and broad frameworks are going to have a meaningful effect on the entire world.[97]

For example, some scholars argue that LAWS can never comply with any legal and ethical standards and that they contravene the so-called Martens Clause of the 1899 Hague Convention on the Laws and Customs of War that "prohibits weapons that are repugnant to the public conscience."[98]

As the head of the Alignment Research Center argues, broad demands such as "Don't build powerful AI systems" are not realistic starting points for serious conversations about how to go about achieving global coordination on AI ethics and regulation. With considerable understatement, he notes that it is a challenging policy problem that requires the sort of geopolitical effort that typically fails even when the stakes are clear and confer notably less pressure to defect.[99]

In a world where it is getting easier for researchers and firms to engage in innovation arbitrage (i.e., innovators and their innovations moving to wherever they receive the most hospitable treatment), some dual-use technologies will be harder to control because they and their creators will gravitate to hospitable countries.

In a world where it is getting easier for researchers and firms to engage in innovation arbitrage (i.e., innovators and their innovations moving to wherever they receive the most hospitable treatment), some dual-use technologies will be harder to control because they and their creators will gravitate to hospitable countries. That is particularly true today because the physicality of technological innovations matters much less than it did in the past.

Moreover, the lines are blurred on exactly what sort of information should be controlled. Treaties seeking to limit the dangers of dual-use technologies could ensnare too much communication and knowledge. As already noted, if not properly targeted and limited in nature and scope, sweeping restrictions on broad classes of technologies could undermine scientific discovery and the many accompanying life-enriching and life-saving benefits that specific algorithmic technologies could bring about.[100] For example, the same advanced AI capabilities that could give rise to killer robots could just as easily give civilization robots that help us with health care or risky jobs. Similarly, a robotic exoskeleton that could equip a soldier with greater warfighting capabilities could also be applied as a life-changing exo-suit to help those with paralysis regain the ability to walk. Society has faced these issues before with the rise of industrial machines and new medical capabilities. There is no escaping these dual-use conundrums, and there is often no easy way of putting the proverbial genie back in the bottle once it is out.

Nonetheless, the risks associated with LAWS are real, and some suggest that other multinational efforts are needed to address them. In a new book on existential risk issues, one of Bostrom's Oxford University colleagues suggests establishing a

[97] Byrne Hobart and Tobias Huber, "Against Safetyism," Pirate Wires, April 17, 2023. https://www.piratewires.com/p/against-safetyism.

[98] Guido Noto La Diega, "The Artificial Conscience of Lethal Autonomous Weapons: Marketing Ruse or Reality?," *Law and the Digital Age* 1 (2018), pp. 1-17. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3607035.

[99] Paul Christiano, "Where I agree and disagree with Eliezer," AI Alignment Forum, June 19, 2022. https://www.alignmentforum.org/posts/CoZhXrhpQxpy9xw9y/where-i-agree-and-disagree-with-eliezer.

[100] Brundage et al., p. 52. https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf.

new “constitution for humanity” to address the broadening scope of existential threats, including AI risks.[101] The U.N.’s Universal Declaration of Human Rights and various other multilateral charters and declarations offer possible templates, but enforcement options are even more limited in this regard and have some notable holdouts, including the United States, which refuses to sign onto the Convention on the Rights of the Child.[102] Additionally, through their seats on the U.N. Security Council, China and Russia can single-handedly hold up progress on any sort of declaration calling for collective action on even the most mundane resolutions, like opposing Russia’s unprovoked war against Ukraine.[103] Adding insult to injury, in March 2023, the U.N. allowed Russia to take over as head of the Security Council despite its continued war against Ukraine.[104] And although critics sensibly decry “the illogic of nuclear escalation,” the threat of mutual destruction has not stopped governments from continuing to spend lavishly on nuclear weapons.[105] In fact, in 2022, Congress approved $51 billion in spending for nuclear weapons with President Joe Biden’s blessing.[106] Meanwhile, Russia recently dropped out of its last remaining nuclear arms control treaty with the United States.[107] Thus, creating new institutions, treaties or declarations focused on AI existential risk likely would not have better outcomes simply by being rebranded as a constitution for humanity.[108]

It is also unlikely that U.N.-led efforts on LAWS control will constrain rogue nations. Again, the U.N.’s history with nuclear arms-control efforts does not bode well for potential LAWS applications. It is often hard to take the organization seriously when it recently allowed a rogue state like North Korea to take over as head of the organization’s Conference on Disarmament, even though, according to the Arms Control Association, the U.N. Security Council “has adopted nine major sanction resolutions on North Korea in response to the country’s nuclear and missile activities since 2006.”[109] Moreover, North Korea withdrew from the nuclear NPT in 2003. Even routine nuclear monitoring efforts often fail. In early 2023, the IAEA reported that 10 drums containing approximately 2.5 tons of natural uranium previously being tracked in Libya had gone missing.[110] If controlling physical weapons or dangerous materials is this challenging, it is hard to imagine how controlling algorithmic systems would be any easier.

If controlling physical weapons or dangerous materials is this challenging, it is hard to imagine how controlling algorithmic systems would be any easier.

[101] Ord, *The Precipice*.

[102] “The United Nations Convention on the Rights of the Child,” Congressional Research Service, July 27, 2015. https://crsreports.congress.gov/product/pdf/R/R40484/25#:~:text=Bush%20Administrations%20played%20significant%20roles,advice%20and%20consent%20to%20ratification.

[103] United Nations, “Russia blocks Security Council action on Ukraine,” United Nations, Feb. 26, 2022. https://news.un.org/en/story/2022/02/1112802.

[104] Julian Borger, “‘Absurdity to a new level’ as Russia takes charge of UN security council,” *The Guardian*, March 31, 2023. https://www.theguardian.com/world/2023/mar/31/absurdity-to-a-new-level-as-russia-takes-charge-of-un-security-council.

[105] Fred Kaplan, “The Illogic of Nuclear Escalation,” *Asterisk*, November 2022. https://asteriskmag.com/issues/1/the-illogic-of-nuclear-escalation.

[106] Ibid.

[107] Vladimir Isachenkov, “Putin signs bill to suspend last nuclear arms pact with US,” AP News, Feb. 28, 2023. https://apnews.com/article/russia-us-nuclear-pact-suspension-ukraine-putin-e579b7562fb816d899e037d1d271a8c5.

[108] Ord, *The Precipice*.

[109] “Amid criticism, North Korea takes over as UN disarmament president,” CNN, June 3, 2022. https://www.cnn.com/2022/06/02/asia/north-korea-un-disarmament-president-intl-hnk/index.html; Kelsey Davenport, “UN Security Council Resolutions on North Korea,” Arms Control Association, January 2022. https://www.armscontrol.org/factsheets/UN-Security-Council-Resolutions-on-North-Korea.

[110] “UN nuclear watchdog says 2.5 tons of uranium missing from Libyan site,” France24, March 16, 2023. https://www.france24.com/en/africa/20230315-un-nuclear-watchdog-says-2-5-tonnes-of-uranium-missing-from-libyan-site.

In sum, no matter how well-intentioned, unenforceable treaties and other amorphous, aspirational agreements are not likely to act as meaningful constraints on the development of powerful algorithmic systems or particular LAWS. Other steps and strategies will be needed.

No matter how well-intentioned, unenforceable treaties and other amorphous, aspirational agreements are not likely to act as meaningful constraints on the development of powerful algorithmic systems or particular LAWS.

## The Uncomfortable Prospect of Bilateral or Unilateral Action

Because of these multilateral enforcement challenges, the danger exists that the same sort of unilateral or bilateral actions we have seen deployed in the past will be used to address certain concerns about LAWS. It is easy to imagine the formation of various coalitions of the willing, or groups of nation-states poised to take action to address serious AI risks or threats posed by other nation-states.[111] An example of this type of cyberwar was the Stuxnet computer virus attack, which was likely a joint effort by the United States and Israel to use computer malware to sabotage the Iranian nuclear program.[112] While the two governments have never acknowledged any involvement in the creation or distribution of the virus that infiltrated and disabled Iran's uranium enrichment facilities, computer security experts almost universally agree that the enormity of the undertaking means that it had to be the product of a nation-state (or multiple nation-states) with sophisticated capabilities in creating such a cyber-weapon.[113]

Whether covert cyber-sabotage was the right move in this case is debatable, especially because the virus also infected many other global systems outside of Iran. On one hand, the attack did seemingly succeed in holding back Iran's efforts to build nuclear military capacity for a period of time. On the other hand, it is unclear whether the Stuxnet effort significantly affected Iranian nuclear enrichment efforts or weapons development, as the country is now evading IAEA nuclear monitoring efforts and recently boasted about developing a hypersonic missile capable of penetrating any air-defense system.[114] Additionally, in early 2023, IAEA inspectors reportedly found traces of near-weapons-grade nuclear material at an underground facility in Iran.[115]

But looking at this example from another angle, it is worth considering whether Israel and the United States could have made the case for more aggressive military action to stop the Iranians directly through some sort of multilateral effort, perhaps via the U.N., instead of opting for covert cyberwar actions. It seems unlikely that the U.N. would have been able to muster enough global support for a broad-based

[111] National Security Commission on Artificial Intelligence, p. 28. https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

[112] Kim Zetter, "An Unprecedented Look at Stuxnet, the World's First Digital Weapon," *Wired*, Nov. 3, 2014. https://www.wired.com/2014/11/countdown-to-zero-day-stuxnet.

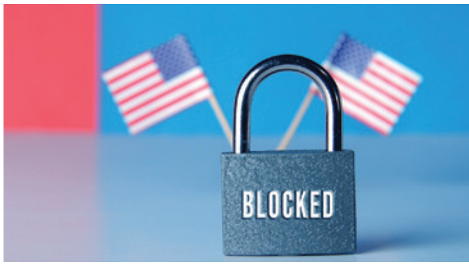
[113] Michael Joseph Gross, "A Declaration of Cyber-War," *Vanity Fair*, March 2, 2011. https://www.vanityfair.com/news/2011/03/stuxnet-201104.

[114] Aresu Eqbali et al., "Iran Says It Has Built Hypersonic Missile," *The Wall Street Journal*, Nov. 10, 2022. https://www.wsj.com/articles/iran-fails-to-provide-answers-to-u-n-in-nuclear-material-probe-11668089209.

[115] Laurence Norman, "U.N. Agency Confirms Iran Produced Enriched Uranium Close to Weapons Grade," *The Wall Street Journal*, Feb. 28, 2023. https://www.wsj.com/articles/u-n-agency-confirms-iran-produced-enriched-uranium-close-to-weapons-grade-7ccd4069.

effort to stop the Iranians from enriching uranium when several other nations had already done the same thing, which is probably why Israel and the United States may have chosen to act secretly. Thus, despite the conundrums of unilateral or bilateral enforcement efforts, because of a lack of effective alternatives, the Stuxnet response could become a template for future state efforts to sabotage the technical AI-related capabilities of rogue actors, however that may be defined.

This possibility is concerning because although cyber espionage or sabotage might stop or slow certain technological capacities of rogue actors, it could also provoke an escalation of hostilities and potentially lead to geopolitical instability through other forms of conflict. Consequently, preemptive military action to deter a threat can be just as risky of a gambit as inaction.



Short of unilateral or bilateral military or covert operations, less risky actions can be taken. Export controls, sanctions, and other trade or investment restrictions will continue to play a role in constraining access to certain technologies that might give rise to existential risks.

Short of unilateral or bilateral military or covert operations, less risky actions can be taken. Export controls, sanctions, and other trade or investment restrictions will continue to play a role in constraining access to certain technologies that might give rise to existential risks, especially goods and services sold to China and other antidemocratic regimes.[116] The United States already imposes a wide array of export controls on various technologies and has been expanding those regulations. In late 2018, the U.S. Department of Commerce's Bureau of Industry and Security announced a "Review of Controls for Certain Emerging Technologies" and launched an inquiry into expanding the list of technologies that would be subjected to the United States' complex export-control regulations.[117]

Of note, many of the technologies under consideration (including AI and robotics) are dual-use in nature, and if restrictive export controls were imposed in a blanket fashion on such technologies, it could seriously undermine U.S. innovation and competitiveness and lead to a loss of companies and talent.[118] Commenting on the effect such rules might have, *The New York Times* suggested that "[o]verly restrictive rules that prevent foreign nationals from working on certain technologies in the United States could also push researchers and companies into other countries."[119] The *Times* also quoted an international trade lawyer who said that if controls were imposed by the United States, "[i]t might be easier for people to just do this stuff in Europe."[120]

For these reasons, export-control mechanisms designed for the industrial era are unlikely to work as well for digital-era issues and, worse yet, overzealous export restrictions could punish U.S. companies more than they might punish foreign

[116] "Mid-Decade Challenges to National Competitiveness," Special Competitive Studies Project, September 2022. , pp. 79-80. https://www.scsp.ai/wp-content/uploads/2022/09/SCSP-Mid-Decade-Challenges-to-National-Competitiveness.pdf.

[117] "Review of Controls for Certain Emerging Technologies" *Federal Register* 83:223 (Nov. 19, 2018), pp. 58201-58202. https://www.govinfo.gov/content/pkg/FR-2018-11-19/pdf/2018-25221.pdf.

[118] Adam Thierer and Jennifer Huddleston Skees, "Emerging Tech Export Controls Run Amok," The Technology Liberation Front, Nov. 28, 2018. https://techliberation.com/2018/11/28/emerging-tech-export-controls-run-amok.

[119] Cade Metz, "Curbs on A.I. Exports? Silicon Valley Fears Losing Its Edge," *The New York Times*, Jan. 1, 2019. https://www.nytimes.com/2019/01/01/technology/artificial-intelligence-export-restrictions.html.

[120] Ibid.

adversaries. This would not only hurt the United States' innovative capacity, but it would also undermine our capacity for building stronger technological offensive and defensive capabilities.[121] As three experts in the field recently concluded, "history suggests that export controls, if not wielded carefully, are a poor tool for today's emerging dual-use technologies. At best, they are one tool in the policymakers' toolbox, and a niche one at that."[122]

While export controls will continue to be a useful strategy in some cases, such restrictions must be narrowly tailored and carefully targeted. The crucial step here lies in identifying which states might have clearly hostile intent and should therefore be cut off.

Thus, while export controls will continue to be a useful strategy in some cases, such restrictions must be narrowly tailored and carefully targeted. The crucial step here lies in identifying which states might have clearly hostile intent and should therefore be cut off—using export controls or other trade and investment restrictions—from engaging in trade in certain narrow classes of AI/ML hardware and software. This is an easier call to make with regard to nations like Iran or North Korea, but it is more difficult to do in a blanket fashion when it comes to China or other countries where U.S. companies do more business. But the Biden administration has recently implemented this strategy with China by imposing restrictions on high-end chips for AI and supercomputing, limiting the country's access to the critical product.[123] In a more tightly integrated trading system with global supply chains and widely dispersed firms and workers, however, such controls could prove difficult to enforce and would be evaded when other sellers emerged. Once again, more broad-based multilateral approaches will likely be required if export controls are going to have a chance of limiting access to such computing capabilities, specific algorithmic or robotic applications.

## Recommendations for Ensuring Continuous Dialogue and Coordination

Continuous Dialogue and Coordination Recommendation

Ongoing communication about these matters can help establish trust and encourage actors to potentially change course or deescalate if certain circumstances arise.

### Prioritize Constant Communication

The most important step for confronting global AI and robotic risk is to keep developers, institutions, and governments thinking and talking about these concerns. Encouraging ongoing and widespread dialogue sounds like a clichéd and unsatisfying solution, but it has great value. Ongoing communication about these matters can help establish trust and encourage actors to potentially change course or deescalate if certain circumstances arise.[124]

Much of this dialogue needs to happen among nation-states, of course. During the Cold War, the United States and the Soviet Union created a direct connection between the White House and the Kremlin—aka, the red telephone (although it was actually a teletype machine)—that was meant to help diffuse tensions and

[121] Loren B. Thompson, "Why U.S. National Security Requires A Robust, Innovative Technology Sector," Lexington Institute, Oct. 8, 2020. https://www.lexingtoninstitute.org/why-u-s-national-security-requires-a-robust-innovative-technology-sector.

[122] Jade Leung et al., "Export Controls in the Age of AI," War on the Rocks, Aug. 28, 2019. https://warontherocks.com/2019/08/export-controls-in-the-age-of-ai.

[123] Gregory C. Allen, "Choking off China's Access to the Future of AI," Center for Strategic & International Studies, Oct. 11, 2022. https://www.csis.org/analysis/choking-chinas-access-future-ai.

[124] Gary Marchant et al., "International Governance of Autonomous Military Robots," *The Columbia Science and Technology Review* XII (2011), pp. 311-313. https://academiccommons.columbia.edu/doi/10.7916/D8TB1HDW.

reduce the risk of war during the nuclear age.[125] It is unclear what the equivalent of that would be in the age of AI, but there will always be value in communication, and nation-states must begin thinking more deliberately about how to discuss AI-related risks when the stakes are so high. One of the best things that could come from ongoing dialogue is the establishment of a better set of international norms for ethical AI development and use. In other words, all roads lead back to soft-law solutions.[126]

Many different nongovernmental international bodies and multinational actors can also play important roles as coordinators of national policies and conveners of ongoing deliberation about various AI risks and concerns. "Direct international regulation is not a realistic option," notes one international governance expert who instead recommends coordinating national law and policy with international action to respond to these types of innovations.[127]

## Take Lessons from Internet Governance: Polycentric Approaches Can Help

Internet management today embodies a polycentric style of governance, with many different actors and governance mechanisms playing a role in ensuring a well-functioning system.[128] Scholars note that arguments in favor of polycentricity include "the notion that it enables governance initiatives to begin having impacts at diverse scales, and that it enables experimentation with diverse policies and approaches, learning from experience and best practices."[129] In other words, polycentricity is just another way of conceptualizing the various decentralized governance ideas and soft-law mechanisms identified throughout this study and previous research.[130]

**Continuous Dialogue and Coordination Recommendation**

**Scholars note that arguments in favor of polycentricity include "the notion that it enables governance initiatives to begin having impacts at diverse scales, and that it enables experimentation with diverse policies and approaches, learning from experience and best practices.**

In the first quarter-century of internet governance, a diverse array of NGOs worked together using ongoing multi-stakeholder negotiations to address a variety of issues. Some of the most important organizations included the Internet Society, the Internet Engineering Task Force (IETF), the Internet Governance Forum, the Internet Architecture Board and the World Wide Web Consortium. These groups worked with governments, industry, civil society groups, university centers and other interested parties to create technical standards for the internet in an iterative, collaborative fashion.[131] The U.N. Internet Governance Forum also works with

[125] Tom Clavin, "There Never Was Such a Thing as a Red Phone in the White House," *Smithsonian Magazine*, June 18, 2013. https://www.smithsonianmag.com/history/there-never-was-such-a-thing-as-a-red-phone-in-the-white-house-1129598.

[126] Kenneth Anderson and Matthew C. Waxman, "Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can," Columbia Law School, 2013. https://scholarship.law.columbia.edu/faculty_scholarship/1803; Cavelty et al.

[127] Kenneth W. Abbott, "An International Framework Agreement on Scientific and Technological Innovation and Regulation," in Gary E. Marchant et al., eds., *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem* (Springer, 2011), pp. 127-128.

[128] John Gerard Ruggie, "Global Governance and 'New Governance Theory': Lessons from Business and Human Rights," *Global Governance* 20 (2014), pp. 8-10. https://scholar.harvard.edu/files/john-ruggie/files/global_governance_and_new_governance_theory.pdf.

[129] Peter Cihon et al., "Should Artificial Intelligence Governance be Centralised?: Design Lessons from History," *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (February 2020), pp. 228-234. https://dl.acm.org/doi/10.1145/3375627.3375857.

[130] Thierer, "Flexible, Pro-Innovation Governance Strategies for Artificial Intelligence." https://www.rstreet.org/research/flexible-pro-innovation-governance-strategies-for-artificial-intelligence.

[131] Thierer, "Soft Law in U.S. ICT Sectors," pp. 90-94.

these organizations to help coordinate governance issues.[132] The International Telecommunication Union and UNESCO also both have an ongoing focus on AI.[133]

Many in the field of internet governance regularly use a phrase made popular by early internet engineers to describe how they kept systems operating through "running code and rough consensus."[134] This idea, which became the unofficial operational motto of the IETF, reflected a pragmatic governance philosophy of continuous iteration and improvement. Perfect agreement on all governance matters was viewed as impossible, but a rough consensus about operational norms became crucial if systems were to grow more robust and reliable. Equally important were the constant tweaks to those systems and the software that powered them.

With AI systems and applications building on top of internet infrastructure and protocols, it is likely that this pragmatic governance philosophy of "running code and rough consensus"—and many of the organizations that make it work—will play a continuing role in overseeing some of the AI-related issues addressed throughout this study. They will join the many other standards bodies, nonprofit organizations and academic institutions already discussed and help them establish governance norms for the AI/ML world in an iterative fashion.

**Continuous Dialogue and Coordination Recommendation**

**Soft-law and less-formal governance mechanisms will have a role to play in the governance of AI risks globally.**

## Embrace Soft Law to Help Establish International Norms

Soft-law and less-formal governance mechanisms will have a role to play in the governance of AI risks globally. For all the reasons identified herein, it remains extremely unlikely that highly centralized or top-down regulatory solutions will work for AI. We are unlikely to witness the development of strict global laws or regulatory bodies for AI or robotics—at least not any with meaningful powers to police the full range of potential risks across nations.

An earlier R Street report identified the various AI "safety by design" ethical frameworks and best practices already being formulated and refined by organizations like the Association of Computing Machinery, the Institute of Electrical and Electronics Engineers (IEEE) and the International Organization for Standardization (ISO).[135] These and other organizations stress the importance of keeping humans in the loop when developing and deploying algorithmic systems to ensure that they are as safe as possible. This is particularly critical for military and law enforcement systems, although experts have noted that a one-size-fits-all solution for meaningful human control does not exist.[136]

[132] "The Internet Governance Forum," Internet Society, last accessed May 7, 2023. https://www.internetsociety.org/events/igf.

[133] "Artificial Intelligence," International Telecommunication Union, last accessed April 7, 2023. https://www.itu.int/en/action/ai/Pages/default.aspx; "Artificial Intelligence," UNESCO, last accessed April 7, 2023. https://en.unesco.org/artificial-intelligence.

[134] A.L. Russell, "'Rough Consensus and Running Code' and the Internet-OSI Standards War," *IEEE Annals of the History of Computing* 28:3 (July-September 2006), pp. 48-61. https://ieeexplore.ieee.org/document/1677461.

[135] Ibid.

[136] Frank Sauer, "Autonomy in Weapons Systems and the Struggle for Regulation," Centre for International Governance Innovation, Nov. 28, 2022. https://www.cigionline.org/articles/autonomy-in-weapons-systems-and-the-struggle-for-regulation.

This is likely why the REAIM conference call to action mentioned earlier concluded by stressing, "[w]e encourage multi-stakeholder dialogue on best practices to guide the development, deployment and use of AI in the military domain to ensure an interdisciplinary discussion throughout of good practices and policies on responsible use of AI in the military domain."[137] This was a tacit acknowledgment by signatories that other governance organizations and mechanisms would play an important role in addressing global AI risks, just as they have with past global risks in the absence of any formal international law or regulatory mechanism.

A variety of soft-law norms and strategies can help address global AI risks. A philosopher at the Naval Postgraduate School identified 11 precepts derived from international laws of armed conflict as well as legal responsibility/liability that he hoped could help classify acceptable and unacceptable practices.[138] For the former category, he identified principles like mission legality, nondelegation of authority and proportional compliance, all which address ethical issues and responsibilities in armed conflict. From the latter category, he identified principles like due care, product liability, and criminal negligence that serve to assign responsibility for harms caused by technological systems or processes.

It is the latter category of norms and best practices where the hard work of professional bodies can help fill in governance gaps. Such multi-stakeholder organizations and standards bodies have global reach and the ability to craft strong but flexible standards for AI oversight and assessment. The question now is how to give current and future soft-law, best-practice frameworks greater visibility and meaning to ensure that they can help shape the development and use of algorithmic systems globally.

## Improve Coordination Among Quangos

Going forward, a greater coordination of soft-law efforts and organizations will be needed if progress is to be made on global AI safety concerns. Scholars have proposed the formation of governance coordinating committees (GCCs) to potentially solve this problem.[139] GCCs represent a type of quango that would help coordinate technological governance efforts among governments, industry, civil society organizations and other interested stakeholders in fast-moving emerging-technology sectors, including AI and robotics.[140] Because it would be impossible for one entity to fully govern any of these rapidly developing, multifaceted fields and the innovations they produce, scholars suggest that GCCs could act as issue managers or orchestra conductors, coordinating a variety of implemented and proposed governance approaches.[141]

**Continuous Dialogue and Coordination Recommendation**

**Because it would be impossible for one entity to fully govern any of these rapidly developing fields, scholars suggest that GCCs could act as issue managers or orchestra conductors, coordinating a variety of implemented and proposed governance approaches.**

[137] "REAIM Call to Action." https://www.mofa.go.jp/mofaj/files/100465549.pdf.

[138] Ibid., p. 333.

[139] Wendell Wallach and Gary Marchant, "Toward the Agile and Comprehensive International Governance of AI and Robotics," *Proceedings of the IEEE* 107:3 (March 2019). https://ieeexplore.ieee.org/document/8662741.

[140] Gary E. Marchant and Wendell Wallach, "Governing the Governance of Emerging Technologies," in Gary E. Marchant et al., eds., *Innovative Governance Models for Emerging Technologies* (Edward Elgar Publishing Limited, 2013), pp. 136-152.

[141] Gary E. Marchant and Wendell Wallach, "Coordinating Technology Governance," *Issues in Science and Technology* XXXI:4 (Summer 2015), pp. 44-45, https://issues.org/coordinating-technology-governance.

GCCs would not be formal regulatory bodies, however. Scholars are instead proposing the creation of a global AI quasi-autonomous nongovernmental organization, or quango. Quangos are NGOs that have a more formal role in the governance of a certain field or technology. Sometimes governments even delegate certain official tasks or responsibilities to quangos that would not usually be carried out by NGOs.

Quangos have been useful in other areas, helping devise solutions to governance-coordination challenges in technically complicated fields. Examples include the IAEA and the International Civil Aviation Organization (ICAO). The IAEA develops global standards on nuclear safety, whereas the ICAO creates standards for international air travel. They are autonomous organizations that work alongside the U.N. to formulate standards and agreements to help create greater trust and security in their respective fields.

The U.N. might be able to create a similar body for AI and robotics. A senior fellow with the Global Center on Cooperative Security recommends that the U.N. form a Global Foresight Observatory for AI and other emerging technologies. It would bring together key stakeholders to deliberate global risk prevention strategies and share information about important developments in this arena.[142] Similarly, two computer scientists have proposed the formation of a new International Agency for AI (IAAI), which would be modeled on the IAEA.[143] This would be a neutral, global nonprofit "with guidance and buy-in from governments, large technology companies, nonprofits, academia and society at large, aimed at collaboratively finding governance and technical solutions to promote safe, secure and peaceful AI technologies."[144]

**Continuous Dialogue and Coordination Recommendation**

**Quangos have been useful in other areas, helping devise solutions to governance-coordination challenges in technically complicated fields.**

Much of the hard work of AI standard-setting and risk management has already been done by professional associations like the IEEE, ISO and ACM.[145] In addition, the Organisation for Economic Co-operation and Development (OECD) has developed a "Framework for the Classification of AI Systems" with the goals of helping "develop a common framework for reporting about AI incidents that facilitates global consistency and interoperability in incident reporting" and advancing "related work on mitigation, compliance and enforcement along the AI system lifecycle, including as it pertains to corporate governance."[146] In the United States, the National Institute of Standards and Technology (NIST) also recently released a comprehensive "Artificial Intelligence Risk Management Framework,"

[142] Eleonore Pauwels, "The New Geopolitics of Converging Risks: The UN and Prevention in the Age of AI," United Nations University Center for Policy Research, April 29, 2019. https://collections.unu.edu/eserv/UNU:7308/PauwelsAIGeopolitics.pdf.

[143] Gary Marcus and Anka Reuel, "The world needs an international agency for artificial intelligence, say two AI experts," *The Economist*, April 18, 2023. https://www.economist.com/by-invitation/2023/04/18/the-world-needs-an-international-agency-for-artificial-intelligence-say-two-ai-experts.

[144] Ibid.

[145] Thierer, "Flexible, Pro-Innovation Governance Strategies for Artificial Intelligence." https://www.rstreet.org/research/flexible-pro-innovation-governance-strategies-for-artificial-intelligence.

[146] "OECD AI Principles overview," OECD.AI, last accessed March 3, 2023. https://oecd.ai/en/ai-principles; "OECD Framework for the Classification of AI Systems," OECD, Feb. 22, 2022, p. 6. https://www.oecd.org/publications/oecd-framework-for-the-classification-of-ai-systems-cb6d9eca-en.htm.

which is a voluntary, consensus-driven guidance document intended to serve as a resource to help organizations in the AI space manage risk and promote trust in AI systems.[147] The NIST framework builds on the ethical frameworks developed by the many different organizations mentioned earlier.

A group of AI researchers and developers argue that "policymakers should collaborate closely with technical researchers to investigate, prevent, and mitigate potential malicious uses of AI."[148] Professional associations with ethical frameworks and standards can facilitate this goal by serving as a baseline for global governance coordination, including the professionalization of flexible AI auditing and impact assessment processes. More formal GCCs could help provide another mechanism whereby AI governance issues are addressed through ongoing collaboration among various parties, both domestically and globally.

**Continuous Dialogue and Coordination Recommendation**

**"Policymakers should collaborate closely with technical researchers to investigate, prevent, and mitigate potential malicious uses of AI."**

In addition, quangos have helped coordinate global governance in other contexts, so an AI quango modeled on the IAEA or ICAO could help craft or even enforce voluntary best practices—or at least offer a forum for ongoing discussions around thorny issues. Perhaps that type of organization could even create a framework for conducting algorithmic audits and impact assessments in a self-regulatory way and then award seals of approval or other awards or certifications for developers following best practices. How such a body would get formed, how membership would work and how the body would be supported financially are all questions that need to be considered.

Again, best practices or codes of conduct for researchers and developers can go a long way toward fostering a culture of responsibility and a greater commitment to safety.[149] Global supply chain management by multinational firms is also a tool for enforcing soft-law norms that have been established through other means. Finally, a variety of transparency laws or other efforts already exist in many national and global governance regimes that can help address global AI risks. In the United States, these include know-your-customer guidelines and whistleblower processes that aim to identify problematic actors in various contexts.[150] Such options should be given a greater chance to help start a conversation about wise technological development and responsible innovation.[151]

National security agencies and defense authorities should also be encouraged to develop ethical AI principles and practices, as the U.S. Department of Defense has

[147] "NIST Risk Management Framework Aims to Improve Trustworthiness of Artificial Intelligence," NIST, Jan. 26, 2023, p. 2. https://www.nist.gov/news-events/news/2023/01/nist-risk-management-framework-aims-improve-trustworthiness-artificial.

[148] Brundage et al., p. 4. https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf.

[149] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014), p. 259.

[150] Bureau of Industry and Security, "Know Your Customer Guidance," U.S. Department of Commerce, last accessed April 12, 2023. https://www.bis.doc.gov/index.php/all-articles/23-compliance-a-training/47-know-your-customer-guidance.

[151] Brian Rappert, "Pacing Science and Technology with Codes of Conduct: Rethinking What Works," in Gary E. Marchant et al., eds., *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem* (Springer, 2011), pp. 109-126.

done.[152] Unfortunately, according to the congressionally chartered National Security Commission on Artificial Intelligence, "[t]here is little evidence that U.S. competitors have equivalent rigorous procedures to ensure their AI-enabled and autonomous weapon systems will be responsibly designed and lawfully used."[153] Hopefully that will change as more efforts like the REAIM conference call to action advance global discussion of AI and robotic risks.

## Fill Gaps with Minilateral Approaches

Minilateral approaches, in which a small coalition of nations work together toward a common goal, will be key in filling the gaps left by the failure of more broad-based multilateral efforts. These types of approaches benefit from "being free of the sclerotic bureaucracy of universal organizations like the UN," and can be more agile and achieve better progress because of the smaller number of participating bodies.[154] Examples of groups that can serve as potential forums for ongoing discussions of global algorithmic risk are the Digital Nations group and the Quadrilateral Security Dialogue, or "the Quad."

**Continuous Dialogue and Coordination Recommendation**

**Minilateral approaches, in which a small coalition of nations work together toward a common goal, will be key in filling the gaps left by the failure of more broad-based multilateral efforts.**

The Digital Nations group was formed in 2014 and now includes 10 member countries (Canada, Denmark, Estonia, Israel, Mexico, New Zealand, Portugal, Korea, United Kingdom and Uruguay). Their goal is to lead by example and become better digital governments more quickly by sharing their expertise with each other on a voluntary and nonbinding basis.[155] One of the group's four primary focus areas is responsible AI use, and they recognize the need to reach a consensus on the best practice for AI use outside of national security and defense.[156] This is a good model for ongoing discussion and coordination on algorithmic governance issues, although it may lack broad enough membership to have a more meaningful impact.

Another example is the Quad, which is an informal effort made up of Australia, India, Japan and the United States.[157] In March 2021, the four countries released "Principles on Technology Design, Development, Governance, and Use" to help their regions and the world move toward "responsible, open, high-standards innovation."[158] The group's effort included the creation of "contact groups on Advanced Communications and Artificial Intelligence focusing on standards-development activities as well as foundational pre-standardization research."[159]

[152] "U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway," DoD Responsible AI Working Council, June 2022. https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artificial-Intelligence-Strategy-and-Implementation-Pathway.PDF.

[153] National Security Commission on Artificial Intelligence, p. 92. https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

[154] Azeem Azhar, *The Exponential Age: How Accelerating Technology Is Transforming Business, Politics, and Society* (Diversion Books, 2021).

[155] Digital Nations, "About," last accessed Aug. 9, 2022. https://www.leadingdigitalgovs.org/about.

[156] "D9 approach for responsible use of AI by Governments," Digital Nations, last accessed April 28, 2023. https://www.leadingdigitalgovs.org/_files/ugd/189d02_8cfd0089d064443fb97c0549b25c77c6.pdf.

[157] Husanjot Chahal et al., "Quad AI: Assessing AI-related Collaboration between the United States, Australia, India, and Japan," Center for Security and Emerging Technology, May 2022. https://cset.georgetown.edu/publication/quad-ai.

[158] "Joint Statement from Quad Leaders," The White House, Sept. 24, 2021. https://www.whitehouse.gov/briefing-room/statements-releases/2021/09/24/joint-statement-from-quad-leaders.

[159] "Fact Sheet: Quad Leaders' Summit," The White House, Sept. 24, 2021. https://www.whitehouse.gov/briefing-room/statementsreleases/2021/09/24/fact-sheet-quad-leaders-summit.

The Quad is viewed by many as an obvious effort to mitigate the risk of Chinese dominance in technological areas, but the effort could also help establish global standards of ethical AI development.[160]

The Quad countries are also part of the Global Partnership on Artificial Intelligence, a broader multi-stakeholder initiative that hosted its first meeting in 2020 to address global AI governance issues in an even more comprehensive fashion.[161] The OECD oversees this effort, which currently includes 25 member states. As with the Digital Nations effort, the goal here is to bring together diverse actors and foster international dialogue and cooperation on best practices that the OECD originally laid out in its 2019 "Recommendation on Artificial Intelligence."[162] The Brookings Institution and the Centre for European Policy Studies also created the Forum for Cooperation on Artificial Intelligence to convene regular AI dialogues among high-level officials from seven governments (Australia, Canada, the EU, Japan, Singapore, the U.K. and the United States) and other AI experts from a broad variety of fields to identify opportunities for international collaboration on AI-related research and development, standards and regulations.[163]

**Continuous Dialogue and Coordination Recommendation**

**One analyst concludes that robotics researchers can help prevent government-based interventions that could hamper AI innovation by creating and championing their own culture of security.**

## Enlist the "Epistemic Community" of AI Developers to Help

AI developers must be "active participants in efforts to monitor the development of systems and technologies with potential military applications," notes one analyst, because they have an interest in ensuring that safer algorithmic systems are widely accepted by policymakers and the public.[164] He concludes that these robotics researchers can help prevent government-based interventions that could hamper AI innovation by creating and championing their own culture of security.[165] Additionally, in many cases, private developers will be the stakeholders who are in the best position to identify new algorithmic risks and work with other developers, security professionals and NGOs to bring those risks to the attention of government actors.[166]

The ultimate question is how to foster constant communication and coordination among what one scholar refers to as the "epistemic community" that is developing around global AI and robotic risks.[167] This notion of an epistemic

[160] Cameron F. Kerry et al., "Strengthening International Cooperation on AI," Brookings, October 2021, p. 39. https://www.brookings.edu/wp-content/uploads/2021/10/Strengthening-International-Cooperation-AI_Oct21.pdf.

[161] Audrey Plonk, "The Global Partnership on AI takes off – at the OECD," OECD.AI Policy Observatory, July 9, 2020. https://oecd.ai/en/wonk/oecd-and-g7-artificial-intelligence-initiatives-side-by-side-for-responsible-ai.

[162] "Recommendation of the Council on Artificial Intelligence," OECD Legal Instruments, May 21, 2019. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[163] "The Forum for Cooperation on Artificial Intelligence," Brookings, last accessed April 4, 2023. https://www.brookings.edu/project/the-forum-for-cooperation-on-artificial-intelligence; Kerry et al. https://www.brookings.edu/wp-content/uploads/2021/10/Strengthening-International-Cooperation-AI_Oct21.pdf.

[164] Geist, p. 320. https://www.tandfonline.com/doi/full/10.1080/00963402.2016.1216672.

[165] Ibid.

[166] Cavelty et al.

[167] Matthijs M. Maas, "How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons," *Contemporary Security Policy* 40:3 (2019), p. 296. https://www.tandfonline.com/doi/full/10.1080/13523260.2019.1576464.

community refers to "a network of individuals or groups with an authoritative claim to policy-relevant knowledge within their domain of expertise [that] share knowledge about the causation of social and physical phenomena in an area for which they have a reputation for competence, and [that] have a common set of normative beliefs about what will benefit human welfare in such a domain."[168] As this report has noted, such an epistemic community has already developed around AI/robotic safety, but it remains quite dispersed, and better coordination will be needed to address current or emerging algorithmic risks.

## Conclusion: Reject Fatalism and Fanaticism When Discussing Global AI Risks

Even if we cannot achieve global consensus on the potential risks associated with particular algorithmic capabilities or successfully craft a formal global regulatory regime to address those concerns, decentralized governance efforts can still help create important operational norms.[169] Governing global AI and robotic risks effectively will require a continuous effort to improve the risk analysis tools we have at our disposal to evaluate current and future threats using the best available knowledge and technology while also tapping the governance mechanisms already in place to seek out constructive solutions to problems as they develop.[170] More than anything else, addressing global existential risks of any variety requires humility, reason and a rejection of fatalism or fanaticism.

Addressing global existential risks of any variety requires humility, reason and a rejection of fatalism or fanaticism.

In many ways, history is repeating itself and we are again witnessing the same sort of fatalistic reasoning that drove extreme thinking and proposals in the early days of the Cold War. During that period, many brilliant, well-meaning intellectuals made dire predictions and extreme proposals for dealing with the existential risk associated with global thermonuclear conflict. In 1951, an influential philosopher predicted, "[t]he end of human life, perhaps of all life on our planet," before the end of the century unless the world unified under "a single government, possessing a monopoly of all the major weapons of war."[171] No global government emerged, yet catastrophe was avoided.

In the AI era, global government solutions are just as unlikely, but there are reasons to believe that the global community can find other ways to work collectively again and muddle through by cobbling together practical governance solutions and encouraging ongoing dialogue about how to best address algorithmic challenges.

[168] Emanuel Adler, "The emergence of cooperation: national epistemic communities and the international evolution of the idea of nuclear arms control," *International Organization* 46:1 (Winter 1992), p. 101. https://www.cambridge.org/core/journals/international-organization/article/abs/emergence-of-cooperation-national-epistemic-communities-and-the-international-evolution-of-the-idea-of-nuclear-arms-control/AD5AB338380EC8691C621B351BC11CE3.

[169] Ingvild Bode and Henrik Hueless, "Autonomous weapons systems and changing norms in international relations," *Review of International Studies* 44:3 (July 2018), pp. 393-413. https://www.cambridge.org/core/journals/review-of-international-studies/article/autonomous-weapons-systems-and-changing-norms-in-international-relations/8E8CC29419AF2EF403EA02ACACFCF223.

[170] Susan R. Dudley et al., "Dynamic Benefit-Cost Analysis for Uncertain Futures," *Journal of Benefit-Cost Analysis* 10:2 (Summer 2019), pp. 206-225. https://www.cambridge.org/core/journals/journal-of-benefit-cost-analysis/article/dynamic-benefitcost-analysis-for-uncertain-futures/F608CEAA98A1AA9337A8DE2AB5881426.

[171] Bertrand Russell, "The Future of Man," *The Atlantic*, March 1951. https://www.theatlantic.com/magazine/archive/1951/03/the-future-of-man/305193.

# KEY TAKEAWAYS

1 Precautionary restraints are most justifiable when the harms are highly probable, tangible, immediate, irreversible, catastrophic, or directly threatening to life and limb in some fashion, but some critics and policymakers define existential risk far too broadly or fail to appreciate how predicting the course of technological developments is severely challenged by knowledge and resource constraints.

2 The most important solutions to technological risk are often more technological innovations to overcome those problems. Blocking future technological innovation and scientific progress will give rise to significant existential risks by depriving society of new technologies that can reduce existing risks and help advance public health and safety.

3 Proposals to impose the global control of AI through a worldwide regulatory authority are both unwise and unlikely to work. Calls for bans or "pauses" on AI developments are largely futile because many nations will not agree to them. No major global power is going to preemptively tie its hands by agreeing to not develop its algorithmic capabilities when adversaries are looking—either overtly or covertly—to advance their own.

4 As with nuclear and chemical weapons in the past, treaties, accords, sanctions, and other multilateral agreements can help address some threats of malicious uses of AI or robotics. Bilateral or unilateral actions may be necessary in certain limited instances when national security threats are clearer and more immediate. But trade-offs are inevitable, and addressing one type of existential risk can sometimes give rise to another, including war.

5 Soft law will play an important role in addressing AI risks. Many different nongovernmental international bodies and multinational actors can play an important role as coordinators of national policies and conveners of ongoing deliberation about various AI risks and concerns.

6 Continuous communication, coordination and cooperation—among countries, developers, professional bodies and other stakeholders—will be essential in heading off risks as they develop and in creating and reinforcing ethical norms and expectations about acceptable uses of algorithmic technologies. In a dynamic, ever-changing technological space like this, new challenges will appear that cannot be envisioned today.

## About the Author

**Adam Thierer** is a senior fellow in the Technology and Innovation Policy program at the R Street Institute in Washington, D.C.