RAND

PERRY WORLD HOUSE
University of Pennsylvania

Expert Insights
PERSPECTIVE ON A TIMELY POLICY ISSUE

CONTRIBUTORS: SARAH KREPS, MILES BRUNDAGE, JAMES D. FEARON, KARL P. MUELLER, JANE VAYNMAN, TRISTAN A. VOLPE

SERIES EDITORS: JIM MITRE, MICHAEL C. HOROWITZ, NATALIA HENRY, EMMA BORDEN, JOEL B. PREDD

The Artificial General Intelligence Race and International Security

September 2025

For more information on this publication, visit **www.rand.org/t/PEA4155-1**.

About RAND

RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Research Integrity

Our mission to help improve policy and decisionmaking through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behavior. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/research-integrity.

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

Published by the RAND Corporation, Santa Monica, Calif.

Cover: Just_Super/Getty Images.

Limited Print and Electronic Distribution Rights

PE-A4155-1

About This Publication

The papers in this publication, commissioned by Perry World House at the University of Pennsylvania and the Geopolitics of AGI Initiative within the RAND Technology and Security Policy Center, explore how competitive dynamics in the pursuit of artificial general intelligence (AGI)—particularly between the United States and China—may shape international security and stability. Drawing on expert perspectives from artificial intelligence (AI) technology, international relations, and national security, the authors contend with whether the greatest risks stem from the ambiguous pre-AGI period or from the rapid, competitive race itself and whether AGI will fundamentally alter the nuclear balance or primarily democratize destructive capabilities. The authors collectively argue that traditional arms control is ill-suited for AGI, proposing instead novel governance models, such as an "AI cartel" to distinguish military from civilian applications. Collectively, the papers highlight strategic dilemmas—speed versus caution, perception versus reality, and competition versus collusion—that demand deliberate choices to ensure that AGI advances international security rather than undermines it.

Perry World House

Perry World House is a center for scholarly inquiry, teaching, research, international exchange, policy engagement, and public outreach on pressing global issues at the University of Pennsylvania. Perry World House's mission is to bring the academic knowledge of the University of Pennsylvania to bear on the world's most pressing global policy challenges and to foster international policy engagement within and beyond the Penn community.

Located in the heart of campus at 38th Street and Locust Walk, Perry World House draws on the expertise of Penn's 12 schools and numerous globally oriented research centers to educate the Penn community and prepare students to be well-informed, contributing global citizens. At the same time, Perry World House connects Penn with leading policy experts from around the world to develop and advance innovative policy proposals.

Through its rich programming, Perry World House facilitates critical conversations about global policy challenges and fosters interdisciplinary research on these topics. It presents workshops and colloquia, welcomes distinguished visitors, and produces content for global audiences and policy leaders, so that the knowledge developed at Penn can make an immediate impact around the world.

Technology and Security Policy Center

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

RAND Geopolitics of AGI Initiative Funding

The Geopolitics of AGI Initiative is independently initiated and conducted within the Technology and Security Policy Center using income from operations and gifts from philanthropic supporters, which have

been made or recommended by DALHAP Investments Ltd., Ergo Impact, Founders Pledge, Charlottes och Fredriks Stiftelse, Good Ventures, Jaan Tallinn, Longview, and Open Philanthropy. A complete list of donors and funders is available at www.rand.org/TASP. RAND donors and grantors have no influence over research findings or recommendations.

Acknowledgments

The editors thank David Glickstein, Gary Cecchine, Paige Smith, Bryan Frederick, and Alexandra Evans for their support in developing this publication.

Summary

As humanity approaches the technological capacity to develop artificial general intelligence (AGI), the race between leading artificial intelligence (AI) powers—particularly the United States and China—is likely to intensify amid broader U.S.-China strategic competition. Perry World House at the University of Pennsylvania and the RAND Geopolitics of AGI Initiative commissioned papers by experts in AI, international relations, and national security to examine the dynamics of the AGI race and its potential implications for international security and stability.

This publication includes papers that explore AGI-related drivers of instability, the race's relationship to the nuclear revolution, and governance challenges and strategies:

- In **"Racing Toward Clarity: How Accelerating AGI Development Could Enhance Strategic Stability," Sarah Kreps** argues that dangers stem largely from the extended period of technological ambiguity before AGI's arrival, increasing risks of miscalculation.
- In **"Unbridled AI Competition Invites Disaster," Miles Brundage** warns that rapid AGI development is driving corner-cutting and threatens to bring about catastrophic accidents, misuse, or violent conflict.
- In **"One Does Not Simply Dismiss the Nuclear Revolution," James D. Fearon** contends that AGI will not overturn mutual nuclear vulnerability but could enable the democratization of weapons of mass destruction and subversion of social cohesion.
- In **"Averting Attacks Against AGI Development: Three Strategic Approaches," Karl P. Mueller** highlights the gap between strategic reality and perception, emphasizing that leaders' beliefs about AGI capabilities shape stability and risks, and explores strategies to manage the AGI race.
- In **"Competition and Collusion: How the AI Arms Race Can Motivate Governance," Jane Vaynman and Tristan A. Volpe** argue that traditional arms control is ill-suited for AGI and propose an "AI cartel" of states and firms to enforce standards distinguishing military from civilian AI systems.

Contents

CHAPTER 1

Introduction: The Impact of the Artificial General Intelligence Race on International Security and Stability

Jim Mitre, Michael C. Horowitz, Natalia Henry, Emma Borden, and Joel B. Predd, editors

Within broader debates about whether, when, and in what form artificial general intelligence (AGI) will arrive lies the critical question: To what extent will competitive dynamics in pursuit of AGI shape international security and stability? Dynamics of the race toward AGI present a new strategic dimension for international politics, particularly because they are unfolding primarily within the context of intensifying U.S.-China strategic competition.

To increase our understanding of AGI race dynamics, Perry World House at the University of Pennsylvania and RAND's Geopolitics of AGI Initiative commissioned papers by experts in artificial intelligence (AI) technology, international relations, and national security to articulate their views on the topic.[1] Each paper comes from a different perspective and offers unique insights on these questions, not only building knowledge to help guide policymakers but also strengthening ties across intellectual communities interested in AI and international security.

Drivers of Instability: Capability Versus Ambiguity

A primary axis of debate among the experts concerns the true source of potential instability from AGI. Does the sheer power of a mature AGI or the murky uncertain period of its development carry the greatest risk of peril?

Sarah Kreps ("Racing Toward Clarity: How Accelerating AGI Development Could Enhance Strategic Stability") posits that the principal danger lies not in AGI's arrival but in whether there is an "extended

[1] When commissioning the papers, the editors did not ask authors to use a particular definition of *AGI*, *racing*, *stability*, or other terms. Authors included their definitions of *AGI* in each piece. Most definitions attribute the following to AGI: a system capable of performing valuable, important, or relevant tasks, at or beyond the levels at which a human being could complete such tasks, across a general (as opposed to narrow) scope, and with a degree of autonomy. Similarly, the papers do not use a single definition, or discuss a single manifestation, of a race, although several authors describe a competition to reach an undefined AGI-enabled weapon capability. In a separate article, Colin Kahl and Jim Mitre articulate five "races" between the United States and China alone (see Colin H. Kahl and Jim Mitre, "The Real AI Race: America Needs More Than Innovation to Compete with China," *Foreign Affairs*, July 9, 2025). Finally, the terms *stability* and *instability* are generally used to refer to conflict between states; however, it is possible that AGI will have profound effects on economic stability, institutional stability, and other conditions in the global system that would not necessarily lead to direct conflict between states.

The papers have been formatted, reviewed for spelling, and lightly edited for consistency and readability but are otherwise presented as submitted.

period of technological ambiguity" that precedes it. During a "pre-AGI zone," states may operate under profound uncertainty about rivals' capabilities and intentions, creating fertile grounds for miscalculation, although states have tools to manage those risks. As a result, Kreps suggests that uncertainty, not capability, will be the primary driver of instability. Paradoxically, a faster, more transparent race toward AGI might enhance strategic stability by creating consistent expectations and a shared understanding of risks. AGI, from this perspective, is not merely a weapon but a "cognitive infrastructure" whose arrival could improve prediction and reduce misunderstandings.

In contrast, **Miles Brundage ("Unbridled AI Competition Invites Disaster")** argues that the unbridled pace of AGI competition is itself the crisis. Fierce rivalries incentivize companies and countries to cut corners with safety and security—a dynamic that he says is already observable in the commercial sector. As AI capabilities advance, the consequences of corner-cutting may also escalate, from deploying flawed consumer products to enabling catastrophic security failures. Brundage forecasts that AI will be capable of most economically valuable computer-based tasks by 2027, which will also dramatically lower barriers for cyber warfare and biological weapon development. These risks, alongside the incentives against implementing careful, methodological regulations, make the speed of AGI racing inherently destabilizing.

The divergence between these two perspectives creates a central strategic dilemma: While Kreps sees speed and transparency as potential antidotes to the poison of ambiguity, Brundage sees speed as a risk factor in itself. Resolving this challenge will likely hinge on whether Brundage is correct about the consequences of competitive pressure and whether competitive pressure makes transparency and shared understanding less likely even after the emergence of AGI.

The Character of the Revolution: Does AGI Repeal the Nuclear Peace?

Besides the path and speed of development, the shadow of the nuclear revolution also plays a central role in the debate over AGI. **James Fearon's ("One Does Not Simply Dismiss the Nuclear Revolution")** analysis is a powerful work of strategic sobriety. He argues that AGI is unlikely to repeal the core logic of the nuclear age: mutual vulnerability to annihilation from nuclear weapons. AGI, he contends, will neither enable a reliable "splendid first strike" against a well-postured nuclear adversary nor create a perfect defense against nuclear attack. The sheer complexity of such tasks, the impossibility of real-world testing, and the range of other options available to a defender (e.g., decoys, randomization, air-gapping, alternate delivery systems) create insurmountable obstacles. Therefore, even a state with massive AGI-driven economic and conventional military advantages cannot safely invade a nuclear-armed peer. The worst-case impacts of AGI, Fearon contends, are therefore potentially overstated.

Fearon posits that, rather than great-power conflict, the primary threat to national security arises from AGI's possible "democratization of [weapons of mass destruction] and mass subversion." For instance, AGI could empower individuals and non-state groups to develop and deploy biological, chemical, and radiological weapons and execute mass subversion campaigns that undermine social cohesion.

Related to Fearon's argument about AGI and the democratization of capabilities, **Karl Mueller ("Averting Attacks Against AGI Development: Three Strategic Approaches")** introduces a crucial distinction between strategic reality and strategic perception. Mueller's paper notes that stability ultimately hinges on what decisionmakers believe and expect about the technology, regardless of its objective state. Even if a disarming, AGI-enabled first strike is a technical fantasy, the fear that a rival is on the cusp of achieving it could motivate preventive action. Brundage also echoes this, warning that a state believing that a rival is close to neutralizing its deterrent may feel an overwhelming "imperative to strike first."

What About Arms Control?

If there are potential risks to stability from AGI, can we count on arms control to solve potential security dilemmas that might emerge? The authors broadly agree that traditional arms control is ill-suited for AGI. **Jane Vaynman and Tristan A. Volpe ("Competition and Collusion: How the AI Arms Race Can Motivate Governance")** provide a sophisticated explanation, arguing that AI exists in an arms control "dead zone." Unlike nuclear technology, which has observable infrastructure in a narrow niche, AI's civilian and military applications are deeply integrated and often functionally identical. As a result, establishing a regulation, monitoring, and verification regime would require intrusive monitoring into states' military and economic secrets.

Vaynman and Volpe thus propose a pivot: Rather than trying to control AI itself, governance efforts should focus on enforcing the distinguishability of its military applications. This enforcement would be achieved not through a universal treaty but through selective collusion: an "AI cartel" of leading states and companies. This cartel would establish and enforce common standards—such as verifiable watermarks, distinct hardware, or specific safety protocols—that make military AI systems observable and distinct from their civilian counterparts.

This solution is a powerful and novel strategy. It shifts the problem from the technically intractable to the politically challenging. Vaynman and Volpe compellingly argue that companies would have rational, commercial incentives to join such a regime, to achieve such potential benefits as capturing market premiums from governments, avoiding exclusion from lucrative markets, and leveraging state-provided resources for compliance. In his paper, Brundage similarly calls for establishing baseline standards, verification, and incentives, for which the cartel model provides a plausible mechanism.

The Problem of Control: Deterrence and Governance in the AGI Era

Given these potentially destabilizing or stabilizing dynamics and the authors' views on arms control possibilities, the authors also explore tools to manage the race to AGI and limit risks. The experts in this publication collectively explore three approaches: deterring attacks, managing technological applications, and creating incentives for restraint.

In particular, Mueller offers a framework that classifies strategies to avert preventive strikes into three categories familiar from deterrence theory. First, *protecting and preempting (denial)* involves implementing defensive measures, such as hardening infrastructure, concealing AGI facilities, and increasing resilience to make attacks appear unlikely to succeed. Second, *threatening and responding (punishment)* relies on the threat of retaliation to make the costs of an attack prohibitively high. However, establishing credible threats will be a central challenge for limited or deniable attacks. Third, *reassuring and rewarding (positive inducements)* involves using positive incentives, such as transparency or technology-sharing agreements, to make the status quo more attractive than attacking. Mueller's three-part framework also highlights the acute challenge of attribution. Although a large-scale kinetic strike would be obvious, many potential attacks (e.g., cyber, sabotage, assassination) could remain covert, undermining the logic of punishment-based deterrence.

Conclusion: Strategic Choices for an Uncertain Future

Collectively, the papers neither accept technological determinism nor offer single, unified predictions and solutions. Instead, they illuminate a set of profound strategic dilemmas that merit careful consideration:

- *The impact of speed on stability.* If there is a nation-state AGI race, would it be more stabilizing if it happens quickly, as Kreps suggests, or should states prioritize caution at the cost of ceding initiative, recognizing the inherent dangers of speed that Brundage identifies?
- *Perceived versus actual capabilities.* Even if Fearon is right that AGI will not upend the nuclear age, how will the perception of AGI's power shape how states behave? Deterrence is a psychological phenomenon, and, as Mueller advises, a belief in new capabilities can be as destabilizing as the capabilities themselves.
- *Contests or cooperation among states and firms.* Is unbridled competition inevitable, or are there opportunities for the selective, collusive governance proposed by Vaynman and Volpe? Vaynman and Volpe state that rivals have a shared interest in preventing the strategic environment from descending into an opaque, unverifiable dead zone.

The task of navigating these dilemmas is formidable for policymakers and private-sector leaders alike. It requires a clear-eyed assessment of the factors driving states and corporations to race toward AGI, a sophisticated understanding of the technology's general-purpose nature, and a willingness to devise incentives that ensure that the path to AGI enhances, rather than undermines, international security.

CHAPTER 2

Racing Toward Clarity: How Accelerating AGI Development Could Enhance Strategic Stability

Sarah Kreps

The prevailing discourse on an AI arms race is steeped in fear and alarmism. According to Kai-Fu Lee, in *AI Superpowers*, "The gap between the global haves and have-nots will widen, with no known path toward closing it. The AI world order will combine winner-take-all economics with an unprecedented concentration of wealth in the hands of a few companies in China and the United States."[1]

Henry Kissinger, Eric Schmidt, and Daniel Huttenlocher applied the tenor of that concern to security: "The shift to AI and AI-assisted weapons and defense systems involves a measure of reliance on—and, in extreme cases, delegation to—an intelligence of considerable analytic potential operating on a fundamentally different experiential paradigm. Such reliance will introduce unknown or poorly understood risks." They conclude that "traditional concepts of defense and deterrence—and the laws of war as a whole—may deteriorate" with implementation.[2]

The formulation holds that accelerating AI, and especially AGI, will inevitably lead to chaos. This paper challenges that premise. I argue that the danger lies not in the race itself but in the extended period of technological ambiguity that precedes AGI's arrival. In this "pre-AGI zone," actors face profound uncertainty about others' capabilities, intentions, and thresholds. Uncertainty rather than capability drives miscalculation. The most dangerous moment in AGI development is therefore less its arrival than the ambiguous interim.

The longer that AGI development remains murky, the greater the risk that states will miscalculate each other's capabilities or intentions. If development happens more quickly and in a way that allows key players to observe progress and align on basic risk models, AGI could actually help stabilize crises by improving prediction, reducing misunderstanding, and creating stable expectations.

Definitions and Assumptions

Evaluating the claim that a fast, transparent move to AGI would increase rather than undermine global security requires clear definitions and assumptions. I use *AGI* to refer to an artificial system capable of autonomous reasoning, learning, and planning across multiple cognitive domains, particularly under conditions of uncertainty, scale, and strategic consequence. This contrasts with narrow AI, which excels at specialized tasks (e.g., target identification) but lacks generalizable cognitive flexibility. AGI functions as a cognitive infrastructure, a system that can simulate, stress-test, and forecast complex global dynamics with far greater

[1] Kai-Fu Lee, *AI Superpowers: China, Silicon Valley, and the New World Order*, Harper Business, 2018, pp. 20–21.

[2] Henry A. Kissinger, Eric Schmidt, and Daniel Huttenlocher, *The Age of AI: And Our Human Future*, Back Bay Books, 2022, p. 21.

fidelity than existing tools, owing to the breadth and diversity of its input processing. It enables large-scale modeling of adversary intent; autonomous decision support in logistics, deterrence, and diplomacy; and strategic cognition at scale. AGI is not an artificial person but a general-purpose engine of prediction and planning with uniquely high utility in geopolitics.

Estimates of AGI's arrival are difficult to project in part because AGI has meant different things to different people. Indeed, the definitional ambiguity is part of the problem. Currently, the world of narrow AGI has proliferated, yielding everything from AI-powered disinformation campaigns to rapid-fire cyberattacks enabled by machine agents to AI-driven surveillance that exacerbates repression and misperception. None of these are AGI, but they can generate strategic ambiguity. I assume that precursors with AGI-like properties (e.g., autonomous multimodal agents, large simulation-capable models) may arrive sooner but that full AGI is further off, a 15- to 30-year window.

The concerns raised in this paper suggest that the question of transparency and visible progress toward AGI is a confounding factor. For the purposes of this argument, I assume a moderate level of transparency, sufficient for key governments and peer companies to observe benchmark performance, training scale, or other signals of capability, even if full architectures or training data remain proprietary. This level of transparency does not mean public open-sourcing or inspection but rather observable indicators (e.g., published evaluations, compute footprints, model disclosures) that allow other actors to track progress and update expectations in a structured way.

How AGI Can Mitigate the Security Dilemma

Robert Jervis, writing on the dynamics of the security dilemma, observed that actions taken by one state to enhance its security can unintentionally make other states feel less secure. At the core of the logic is ambiguity, the uncertainty about the other side's capabilities, intentions, or thresholds. When states cannot distinguish defensive measures from offensive ones, they assume the worst. This perception triggers a feedback loop in which each side responds by expanding its capabilities, leading to arms racing and the fear that waiting could mean facing a stronger adversary later.[3] Under such conditions, it can become strategically rational to strike sooner, rather than risk falling behind, but at the least can lead to standoffs, crises, and escalation. The early nuclear era shows how this dynamic played out. From 1945 to 1962, the United States and the Soviet Union operated under deep uncertainty, which led to repeated crises in Berlin, Korea, and Cuba, along with several near misses between nuclear-armed rivals.

According to Jervis, new technologies exacerbate the security dilemma's key factor of uncertainty. Without knowing whether a new capability is defensive or offensive, states will assume the worst. In the AI domain, the security dilemma intensifies because states cannot easily observe or verify one another's capabilities. Unlike nuclear systems, AI lacks visible indicators, such as missile counts or deployments. Developers embed meaningful progress in datasets, proprietary code, and private compute clusters, making it difficult for outsiders to assess. That commercial developers now lead much of the most advanced AI development and often pursue timelines and incentives that can diverge from those of national governments means a decentralized structure whose consequences could weaken signaling, impair coordination, and increase the chance of escalation sparked not by deliberate policy but by misaligned corporate actions or competitive missteps.

But if ambiguity drives instability, the move to AGI could actually mitigate the security dilemma. Much of the current discourse frames AGI as a military force multiplier, specifically an autonomous system that enables faster targeting, quicker escalation, or more-lethal outcomes. That vision misunderstands the real

[3] Robert Jervis, "Cooperation Under the Security Dilemma," *World Politics*, Vol. 30, No. 2, January 1978.

danger. What destabilizes great-power relations is not just speed or power, but uncertainty about how rivals interpret risk, how they respond to pressure, and where they draw their redlines.

AGI could help address this ambiguity by acting as cognitive infrastructure, a system that enables modeling the complexity of crises across domains. For example, a well-designed AGI system could simulate the economic blowback of sanctions, trace the domestic consequences of a blockade, or model the likelihood of escalation through proxies. Rather than acquiring targets or controlling weapons, AGI would forecast ripple effects, alert leaders to second- and third-order consequences, and bring to the surface the hidden risks embedded in their own strategies. By doing so, AGI would function as a diagnostic tool layered above the military and diplomatic systems it helps contextualize.

Most crises escalate not because adversaries want war but because they misread each other's signals, intentions, or thresholds. AGI could narrow that margin of error by providing a shared simulation environment that helps decisionmakers anticipate how different actors might interpret a given action or doctrine. Indeed, one criticism of AI proliferation in the media or creative world is the homogenization of content, but such flattening becomes a virtue rather than a vice in an AGI world on the battlefield. Even partial convergence around the likely consequences of specific maneuvers could anchor diplomacy in shared, probabilistic baselines. As escalation timelines shrink and new domains like cyber, space, and information disrupt traditional deterrence models, aligning forecasts becomes more important than aligning intentions. AGI could help close that gap by enabling shared simulation and strategic foresight.

Realizing the potentially stabilizing features of AGI requires an institutional model grounded in political and technological realities. A centralized, supranational AGI under United Nations control is both politically and technically implausible. A more realistic model involves a distributed set of independently developed simulation platforms. These systems, built by governments, commercial labs, or coalitions of the two, could adopt shared protocols for stress-testing, input assumptions, and structured output formats without requiring full openness. Participation would be voluntary and driven by self-interest. States and firms could contribute to simulations without giving up sovereign control or proprietary systems. This kind of decentralized structure offers a flexible alternative to rigid, top-down agreements.

The incentives for transparency may be asymmetric, to be sure, and are conceptually different from arms control antecedents. Nuclear systems were hardware-bound, visible, and countable, ideal conditions for treaty-based verification. AGI is software-defined, fluid in its capabilities, and unlike missiles or fissile material, code can be copied, concealed, or selectively disclosed. This opacity raises a genuine prisoner's dilemma: If one state commits to transparency and another defects, the transparent actor may find itself at a strategic disadvantage.

Democracies may see value in open benchmarking to signal restraint or build alliances, whereas authoritarian states, or any actor betting on a temporary strategic edge, may resist exposure. China, for example, might calculate that secrecy offers a first-mover advantage in crisis modeling or strategic deception. Yet even rivals have reason to avoid catastrophic miscalculation. In this sense, transparency could be incremental rather than binary, a series of verifiable steps that reduce the probability of worst-case escalation. Still, different states interpret risks through divergent institutional, cultural, and doctrinal lenses. AGI could help bridge these gaps by standardizing how actors model escalation and assess consequences.

What destabilizes great-power relations is not just speed or power but uncertainty about how rivals interpret risk, how they respond to pressure, and where they draw their redlines.

Engaging the Counterargument

One counterargument might be that even rapid AGI development could lead to catastrophic outcomes. One scenario could involve a slower, fragmented trajectory that produces a patchwork of unregulated, opaque "almost-AGIs" with unpredictable failure modes. These systems, developed in secret and without shared standards, would deepen strategic ambiguity and fuel arms race dynamics. Worse, even if some AGI systems were shared, they could still compound flawed assumptions, distort decisionmaking, or amplify existing misperceptions if governments rely on them too heavily or treat outputs as neutral truth. Even a seemingly transparent AGI ecosystem, according to this logic, could reinforce rather than reduce instability. Another concern might be that while hallucinations are rare, they could still produce unpredictable outputs that lead to inadvertent escalation. Hallucinations, errors caused not by faulty inputs but by the probabilistic nature of machine learning models that rely on statistical associations rather than fixed logic, are becoming less frequent. Still, even in an AGI context, such errors could persist, and if trusted and allowed to compound across multiple steps in a decisionmaking chain, they could unintentionally drive escalation.

Indeed, while model reliability is improving, the risks from both strategic ambiguity and residual model failures remain. Acceleration, if paired with structured transparency, could help manage both. Compute tracking, open hardware audits, and energy-use signatures can verify development activity and allow states or firms to signal capabilities and restraint without intrusive inspection. The Trump administration's AI Action Plan, by endorsing open-weight and open-source models, supports this kind of verifiability and signaling.[4] It promotes a shared technical baseline that enables independent auditing and more-predictable norms of deployment, which can help avoid misinterpretation and crisis escalation.

In the case of hallucinations, the challenge is to develop clear, transparent, and shared protocols for how to respond when such misfires occur, however rarely. Research on user interaction with complex AI systems remains limited. As AGI systems diffuse, decisionmakers will inevitably interact with them, meaning that the sooner that militaries can test system fidelity and develop protocols for addressing known failure modes, much as they do with aircraft maintenance records, the less uncertainty and the less room for escalation.

Conclusion—Racing Toward Stability

Despite the doom-laden rhetoric, AGI need not be destabilizing. Indeed, in this paper, I have argued that moving quickly but thoughtfully might actually be safer than dragging development out in a fog of uncertainty. The key is to approach AGI as shared infrastructure rather than as a secret weapon for whoever gets there first, which is how the race is currently cast.

AGI will indeed bring serious risks. But keeping development shrouded in secrecy and pretending that countries are not making rapid progress creates its own dangers. The longer that militaries stay stuck in this limbo in which everyone knows that transformative AI is coming but no one wants to plan for it openly, the more likely we are to stumble into exactly the kind of crisis that well-designed AGI could help navigate.

---

[4] White House, *Winning the Race: America's AI Action Plan*, July 2025.

CHAPTER 3

Unbridled AI Competition Invites Disaster

Miles Brundage

Competition between companies and countries to develop and deploy AI has many benefits, but as with many products and services, unbridled competition carries risks, such as corner-cutting on safety. The stakes of unbridled competition are particularly severe for a technology that exceeds human intelligence and which is seen by the most powerful public and private actors in the world as essential to their future.

Because AI is inherently challenging to build safely and securely and there are benefits to being first, corners will be cut to gain an advantage, and the sheer pace and complexity of change will make international crises likely. Corner-cutting is already happening today as a result of fierce competition between companies, and the consequences of unbridled competition will grow as AI capabilities rapidly improve in the next few years and as the technology moves from being primarily applied to relatively lower-stakes commercial applications to being adopted at scale for national security use cases.

Catastrophic outcomes can still be prevented, but only if vigorous action is taken soon to ensure that rigorous safety and security standards are adopted and verified globally.

AI Will Progress Quickly

The global AI community, led by a handful of U.S. companies, has discovered two key insights over the past decade. Both indicate that AI will continue to progress rapidly in the next few years.

First, the most effective way to build AI systems is to use a large amount of computing power and data to train large neural networks, rather than manually coding in human knowledge. AI pioneer Richard Sutton described this as the "bitter lesson" because it means that many of the ingenious ideas that computer scientists have come up with ultimately become obsolete when you have sufficiently powerful computers and enough data. Each year since Sutton's 2019 essay has provided more evidence for this perspective.[1]

Second, there is copious room to continue scaling the latest iteration of this paradigm, and even the very early fruits of that paradigm rival human intelligence in many economically important respects. In this latest paradigm, *large language models* first learn language and world knowledge by ingesting large swaths of the internet and are then trained with *reinforcement learning* (trial and error) to solve harder and harder problems.

By allowing AI systems to learn how to reason through their own experiences rather than from humans telling them how to do so, this paradigm has the potential to significantly exceed human intelligence in any domain where a source of feedback is available, just as AlphaGo exceeded the best humans at the game of Go. Despite being very early in its trajectory, this new scaling direction has borne incredible results already and is transforming many desk jobs, such as computer programming.

[1] Rich Sutton, "The Bitter Lesson," webpage, Incompleteideas.net, March 13, 2019.

My argument here does not depend on concepts like AGI (artificial general intelligence) or ASI (artificial superintelligence), which mean so many things to so many people that their usefulness is sometimes unclear. But to situate my perspective relative to others, I will predict that by almost any reasonable definition one might use for AGI or ASI, and absent significant efforts to prevent this outcome, we will at least achieve the "digital-only" version of AGI and ASI by the end of 2027 (by digital-only, I mean that I take no position on the pace of progress in robotics). Put differently, I predict that by the end of 2027, almost every economically valuable task that can be done *on* a computer will be done more effectively and more cheaply *by* computers.

There will be unevenness in the pace of change—for example, AI will likely be superhuman in almost all aspects of coding and math by the end of 2025, let alone 2026 or 2027. Additionally, just because something can be done does not mean that it will be done (or done at scale), and long after 2027, there may still be economic demand for a human in the loop for various purposes.[2] But as the cost of AI capabilities steadily decreases, humans' primary economic value-add will be their humanness (including the mere fact of being human, as well as their specific identity) rather than their raw intelligence.

These AI capabilities, and the speed with which they will arrive, will be destabilizing by default because they will simultaneously disrupt many aspects of the economy, domestic politics, and international security and because AI development and deployment require great care in order to be done safely and securely.

It Is Hard to Make AI Safe and Secure

Given the trajectory described above, it is natural for companies and countries to compete to gain an advantage in developing and deploying AI. But doing this without imposing severe risks on oneself and others is easier said than done.

AI is definitionally dual use, since it is nothing more or less than the ability to solve a very wide range of problems with software. But the bitter lesson–inspired approach to AI scaling adds risks. The data going into these systems are far beyond a scale at which humans can carefully review them, researchers are struggling to make sense of the inner workings of the systems, and the latest paradigm (reinforcement learning) gives AI systems incentives that do not always align with human interests. For example, current AI systems sometimes produce what they expect humans will *believe* is a solution to their coding problem rather than actually solving the problem at hand. AI systems also readily pander to human biases. In short, we know how to build ever-smarter systems, but we do not yet know how to understand or control them reliably.

Even assuming perfect control by the users of these systems, the potential for malicious use and societal disruption is substantial. Today, the strongest argument for an AI system not being particularly dangerous is that it is not smart enough to do much harm, even if it (or its user) tried to cause harm. Given the pace of progress, this argument (an "inability argument")[3] will not last much longer, and the focus will shift to how effective the safety and security mitigations are. By the end of 2025, we can expect AI to enable significantly more-autonomous systems that have more widely appreciated economic impacts, dramatic reductions in the difficulty of carrying out biological weapon attacks, and a transformation of cybersecurity.

Security is also difficult, and companies at the frontier are far from ready to stop sophisticated state attackers from stealing their intellectual property and using those stolen frontier AI capabilities in more

[2] Rebecca Crootof, Margot E. Kaminski, and W. Nicholson Price II, "Humans in the Loop," *Vanderbilt Law Review*, Vol. 76, No. 2, May 2023.

[3] Joshua Clymer, Nick Gabrieli, David Krueger, and Thomas Larsen, "Safety Cases: How to Justify the Safety of Advanced AI Systems," arXiv, arXiv:2403.10462, March 18, 2024.

reckless or malicious ways.[4] Thus, even if the United States remains in the lead, other states will not be far behind, by default, and will primarily be limited by the computing power they have available for adapting and running AI systems.

Corner-Cutting Is Tempting and Common by Default

The idea that companies will struggle to maintain high safety and security standards amid a competitive environment is not a controversial claim. It is an explicit assumption in the policies of the three leading AI companies (OpenAI, Anthropic, and DeepMind), each of which states that they may relax their safeguards if others do so.[5]

Corner-cutting is also already evident in company behavior. As two examples, consider the timing of model deployments and information security.

System cards published by frontier AI companies typically discuss timelines for external testing that range from a small number of days to a small number of weeks. This compares unfavorably with what was more common two years ago, when companies sometimes "sat on" models for a period of many months. Some of this acceleration can be explained by maturation of safety processes, but some can also be explained by a tighter competitive environment. If the maturation of safety processes fully explained the acceleration, then we would see a greater fraction of known issues (or issues that could have been easily known) being resolved prior to deployment. But in fact, we regularly see cases like the rollback of OpenAI's sycophantic variant of GPT-4o and reports of cheating on coding problems for Anthropic's Claude Code and other widely available AI coding systems.

Regarding information security, independent experts at RAND have found that frontier companies are years away from reliably defending themselves effectively against sophisticated state attackers, a timeline that compares unfavorably with these companies' stated expectations of building extremely capable systems in a matter of months to years.[6] OpenAI was reportedly hacked in 2023, a Chinese national reportedly stole trade secrets from Google in 2024, and Anthropic only recently claimed to have achieved protection against sophisticated *non-state* attackers.[7]

Crises Are Likely

I assume that the private sector will continue to lead AI capability development but with increasing support from the government (e.g., via security assistance, subsidies, and expedited building of data centers) and increasing use of advanced AI by governments for a range of national security applications. I also assume that at least some government agencies will possess *frontier AI* systems of their own (that is, systems that are

[4] Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott, *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*, RAND Corporation, RR-A2849-1, 2024.

[5] OpenAI, "Our Updated Preparedness Framework," webpage, April 15, 2025; Anthropic, "Anthropic's Responsible Scaling Policy," webpage, last updated May 14, 2025a; Allan Dafoe, Anca Dragan, Four Flynn, Helen King, Tom Lue, Lewis Ho, and Rohin Shah, "Updating the Frontier Safety Framework," webpage, DeepMind, February 4, 2025.

[6] Nevo et al., 2024.

[7] Cade Metz, "A Hacker Stole OpenAI Secrets, Raising Fears That China Could, Too," *New York Times*, July 4, 2024; U.S. Department of Justice, "Chinese National Residing in California Arrested for Theft of Artificial Intelligence-Related Trade Secrets from Google," press release, last updated February 6, 2025; Anthropic, "Activating AI Safety Level 3 Protections," May 22, 2025b.

AI-related security crises could take several forms, including a flash war, a preemptive war, a state or non-state actor using AI to carry out a catastrophic biological weapon attack or cyberattack, and loss of human control over a frontier AI system

among the most capable and expensive), though these will typically be obtained from the private sector.[8] In some cases, private-sector and public-sector interests will be aligned, and in other cases, they will come apart. For the purposes of simplicity, I assume that companies will generally have cooperative relations with their host governments, amid competition between companies and between countries. Relaxing this assumption could make additional classes of crisis possible.

Bureaucratic procurement processes, risk aversion, and a strong bias toward human control may slow the integration of AI in national security contexts, but the trend line is clear. The costs of delay and applying disproportionate caution (compared with other states) will grow intolerable due to competition, and warfare—like bureaucratic processes in the private and public sectors more generally—will be increasingly automated, making various kinds of crises likely.

AI-related security crises could take several forms, including a flash war, a preemptive war, a state or non-state actor using AI to carry out a catastrophic biological weapon attack or cyberattack, and loss of human control over a frontier AI system. Each of these is quite different, but they share the common theme of being made more likely by fierce competition and fast-moving technological developments. The first two involve action by militaries, and the last two do not necessarily, but in each case, the threat could still arise regardless of whether the underlying AI capabilities emerge from the public or private sector.

A *flash war*, in this context, is the military analogue of the stock market "flash crash" of 2010. It is a war that occurs in a flash, in the sense that it emerges from rapid interaction between automated systems that cause significant harm before humans can make sense of and intervene to stop the escalation. Such a conflict could erupt and intensify at speeds that outpace human cognitive and decisionmaking capabilities, leading to devastating outcomes that neither side initially intended or desired.

The core danger here lies in automated systems, or human operators overly reliant on AI-driven recommendations, reacting to perceived threats or provocations with machine-speed responses, creating a cascading chain reaction of offensive and defensive actions before human leaders can intervene, de-escalate, or even fully comprehend the unfolding situation.

Traditional arguments against the likelihood of rapid, unintended escalation are significantly weakened in an AI-suffused security environment. Historically, even in tense crises, there has been time, however brief, for human leaders to communicate, assess intentions, and make deliberate choices. AI-driven flash wars could bypass these human loops entirely or present decisionmakers with scenarios that have already escalated beyond manageable thresholds. Hotlines and established diplomatic channels may also prove too slow without investments having been made in advance to prepare for AI-related flash wars, and there is no public evidence yet that much groundwork has been laid for preventing such risks. AI systems also may not be deterred from escalation in the same way that humans are.

Rapid progress in AI may also precipitate a *preemptive war*, in which a state initiates military action to prevent another state from achieving or deploying AI capabilities perceived as constituting an intolerable future threat. AI capabilities could be expected to, e.g., render the attacker's own defenses obsolete, enable an

[8] Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety," arXiv, arXiv: 2307.03718, November 7, 2023.

overwhelming future offensive, or achieve a level of general capability that fundamentally alters the global balance of power through a massive gap in economic growth rates.

The dual-use nature of AI, driven by intense commercial and national security competition, could fuel worst-case assumptions about adversaries' progress and ultimate intentions. Indeed, a similar dynamic has led to a proliferation of competing AI companies that often view each other with suspicion and are sometimes culturally defined in opposition to one another (e.g., with some companies viewing themselves as safer, less ideological, or more transparent versions of one another). If a state believes that another is close to developing AI that could, for example, instantly neutralize its nuclear deterrent or provide an unstoppable offensive cyber capability, the perceived imperative to strike first, even with incomplete information, could become overwhelming.

The usual arguments for why preemptive war is unlikely are also weaker in the AI context. A somewhat stable equilibrium was ultimately reached during the Cold War after a long gestation period for strategic concepts like mutually assured destruction, substantial investments in diplomacy and military-military dialogue, and the development of confidence-building measures. In contrast, the strategic situation with AI could be dramatically transformed within weeks or months, as it regularly is in the private sector. The sheer pace of AI progress means that the time frame for such a point of no return could be, or be perceived as, very short.

A *non-state attack*, in this context, refers to the use of advanced AI capabilities by actors other than recognized nation-states—such as terrorist organizations, extremist groups, sophisticated criminal enterprises, or even well-resourced individuals—to inflict significant harm. This could involve, e.g., AI-assisted development and deployment of novel biological or chemical agents, highly autonomous and scalable cyberattacks, sophisticated disinformation campaigns leveraging deepfakes and personalized propaganda, or the orchestration of attacks using swarms of AI-guided drones.

As it achieves competence in biological reasoning, computer programming, and other areas, AI lowers the barrier to entry for sophisticated attacks. Indeed, leading companies have noted that this is beginning to occur (though the trend has not been fully realized and will be much further along by the end of 2025).

Traditional arguments against the capacity of non-state actors to execute high-impact, technologically advanced attacks often center on their limited resources, lack of state-level infrastructure, and inability to access or weaponize cutting-edge science and technology. AI challenges these assumptions directly. Access to a highly capable AI could substitute for vast human capital and expensive research and development (R&D) infrastructure, allowing a small group or even an individual to achieve outcomes previously thought to be the preserve of state programs. While states might maintain an edge in the absolute frontier of AI development, an AI model that is merely "good enough" to design a novel pathogen or execute a crippling cyberattack will likely become accessible much faster than effective global safeguards can be implemented, at least on the current trajectory.

Finally, a *loss-of-control event* refers to a scenario where advanced AI systems, either individually or collectively, begin to operate in ways that are misaligned with human intentions and values, and humans lose the effective ability to oversee, correct, or shut down these systems, leading to significant and potentially catastrophic unintended consequences.

Given the strong language understanding and growing capabilities of AI systems, it is becoming easier and easier to create systems that know what humans want in a particular situation. Yet, perhaps counterintuitively, it is harder to ensure that systems care what humans want. And for many possible goals that AI systems may be imbued with (e.g., maximizing mathematical understanding), they may pursue dangerous subgoals along the way, such as *self-exfiltration*, in which an AI "hacks itself out" of a computer network, in order to pursue its higher-level goals with less interference. While there is reason to hope that there are solutions to all

of these problems, comprehensive solutions are not yet known, and they may require great care (and significant trade-offs) in order to be implemented effectively.

Aligning Interests for a Better Path Forward

The individual incentives facing companies and countries unfortunately point toward corner-cutting in AI by default. But countries also have a higher-level incentive—if a poorly perceived one today—to try to escape from this dangerous spiral, assuming that the costs of making that escape are manageable and if the risks of not doing so are sufficiently large.

Whether we avoid AI crises will ultimately come down to whether we can sufficiently invest in three foundational areas of AI governance: articulation of baseline safety and security standards, auditing and verification of compliance with standards in a way that respects each party's expectations of security and sovereignty, and incentivizing participation in this regime. Progress in these areas is slower today than progress in AI capabilities. But effective AI governance could be dramatically accelerated with a concerted effort and would cost a trivial amount compared with the current build-out in AI capabilities, as well as the costs of the crises described above.

Baseline Safety and Security Standards

Establishing common safety and security standards for AI is a critical step toward avoiding the outcomes above. It is not sufficient on its own, because such standards can be disregarded if one party decides that its interest in pursuing an advantage is worth the cost, which we consider in the next two subsections. But without an AI analogue to norms against nuclear proliferation and in favor of robust command and control, it will be difficult to get efforts to avoid crises off the ground. Fortunately, common ground is likely discoverable in three areas:

- *Maintaining human oversight of AI:* While there will and should be healthy debate about what kinds of tasks should be delegated to AI, and about the level of abstraction at which humans should be involved in AI decisionmaking, those are mere details compared with the worst-case scenarios that AI researchers worry about. Neither the United States nor China has an interest in the other losing control of its AI systems entirely, and early steps have been taken to ensure a norm of human judgment over the use of force.[9]
- *Information security at the frontier:* Avoiding unintended proliferation of or tampering with the most-capable AI systems is foundational to ensure that there is a well-defined set of players who can align on a certain set of best practices and to prevent rogue actors from stealing and then misusing advanced AI. While companies and countries have an interest in advancing their relative positions in AI, and this will sometimes incentivize hacking one another, there should be room for agreement on some floor for security at some level of AI capability (e.g., for the five to ten most capable projects). The United States and China have an interest in avoiding unintentional proliferation, via theft, to non-state actors and smaller states, and they might even have an interest in both sides having robust security against one another. By analogy, it is easy to imagine a few edge cases in which the United States or China would want to com-

[9] Michael C. Horowitz, "Autonomous Weapon Systems: No Human-in-the-Loop Required, and Other Myths Dispelled," *War on the Rocks*, May 22, 2025; Jarrett Renshaw and Trevor Hunnicutt, "Biden, Xi Agree That Humans, Not AI, Should Control Nuclear Arms," Reuters, November 16, 2024; Bureau of Arms Control, Deterrence, and Stability, "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," U.S. Department of State, last updated November 27, 2024.

promise the other's nuclear command and control, but on the whole, each would rather the other country have robust nuclear security than weak nuclear security. AI security may be a sufficiently difficult problem (e.g., requiring robust stress-testing of secure chip designs) that the choice is in fact between a joint push for high security, with designs stress-tested in the scientific literature, and shared insecurity.

- *Extending chemical, biological, radiological, and nuclear nonproliferation to AI outputs:* In the long term, AI systems that are widely available will be capable of intellectual feats that are jealously protected today and would be catastrophic if unleashed on a society with as much exposure to, e.g., pandemics as we face today. But there will be an intermediate period in which only the very most capable AIs are significant game changers for malicious actors, and it matters a lot which safeguards are applied to closed models and which models are open-sourced. Standards could be devised to extend this period while society improves its resilience to AI misuse. For example, there could be standards for the rigor of "marginal risk" assessments for open-source models. And there could be a norm for the level of rigor expected of independent safety testing for closed models—e.g., that the effort required to elicit certain dangerous capabilities from an AI system must be greater, when measured in dollars, than it would cost to create that capability through other means, or a notification period will be provided before release. Notably, treaties on chemical and biological weapons have not deterred powerful (and even some more minor) states from developing related capabilities—e.g., the Soviet Union blatantly violated the Biological Weapons Convention. However, for reasons discussed below, there is more hope for nonproliferation of frontier AI models than biological insights, and it does seem that norms against biological weapons have at least slowed proliferation to minor states and non-state actors.

The existence of common ground on its own will not be sufficient to establish norms in these areas—indeed, nuclear hotlines were not developed until the Cuban Missile Crisis dramatically underscored the risks of nuclear miscalculation. The task before us is to articulate and act on this common ground before a crisis occurs, which also requires proactively finding ways to assess and incentivize compliance with common standards.

Assessing Compliance

Suppose two or more companies or countries want to follow the rules above, but they worry that the other party will cheat. This is why a verification regime would be needed, as it was with nuclear weapons and chemical weapons.

Frontier AI development currently exhibits a critical dependency: vast, identifiable, and trackable computational resources, commonly referred to as *compute*. The supply chain for specialized AI hardware (e.g., advanced graphics processing units and tensor processing units) is highly concentrated, and the construction of large-scale training clusters involves significant, observable infrastructure investment and energy consumption. These choke points could offer valuable insights into capabilities and development trajectories, forming a key pillar of a verification system modeled after the International Atomic Energy Agency (IAEA) and the Organisation for the Prohibition of Chemical Weapons.[10]

As a simplified summary of a complex supply chain, today, it is U.S. companies, like NVIDIA, working in partnership with TSMC in Taiwan, Samsung in Korea, and ASML in the Netherlands, that produce the lion's share of advanced AI compute. Over time, China may produce a larger fraction of global computing power domestically. A bifurcation in the supply chain could introduce additional complexities to tracking

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

global compute and verifying how it is used—e.g., if there are differences in the technologies used in the two supply chains that make it harder to carry over lessons from one to the other or if reduced interdependence reduces incentives to negotiate agreements. But compute-based governance could remain feasible as long as leading-edge semiconductor manufacturing continues to be extremely capital intensive (and thus tractably observable) wherever it occurs.

By default, there will be an intermediate degree of transparency about AI. Companies and countries will be able to make, and currently make, informed estimates about the approximate scale of computing power that others possess, and steps can be taken to reduce uncertainty (e.g., via more-rigorous export control enforcement internally and transparency or reporting requirements domestically).

Given that computing hardware is foundational to AI capabilities and will remain a key differentiator even as AI gets more capable, this is a promising start.[11] However, even exact knowledge of where computing hardware resides would not be sufficient to ensure strategic stability or convergence on a shared set of norms, for a few reasons. First, there can be (and already is) obfuscation of who actually controls a given quantity of computing power (e.g., one country might operate a data center that is located in a different country).[12] Second, computing power can be used in many different ways, and it is difficult (though not necessarily infeasible, as discussed below) to verify how it is actually being used. And third, the hardware and software of AI are important, but so are the algorithmic insights used to train a given AI model on a given piece of hardware. These algorithmic insights allow an actor to get more "bang for their buck" with a given amount of compute (though it is still preferable to have more compute rather than less).

The core arms control challenge (in general and in the context of AI) lies in achieving sufficient transparency and access to ascertain compliance without compromising legitimate security interests or intellectual property.[13] There is ongoing work in academia and industry on *AI auditing*, where independent parties verify the properties of an AI system. Methods developed for conducting such audits with a high degree of rigor could—if significantly accelerated and paired with diplomacy—build a bridge toward international verification of agreements on AI. Such research investments would ultimately pay for themselves in the stability they enable. The more effectively these techniques are designed in a way that avoids leaking sensitive information not related to noncompliance, the more politically palatable it will be to deploy them in high-stakes contexts and the more valuable they will be as early warning signals of an intent to violate safety and security standards. In the initial stages, and as a later-stage supplement to more-intrusive verification, confidence-building measures can also play a vital role.[14] Efforts to accelerate progress in these areas could include joint research projects on AI safety and control, dialogues between AI safety teams from different companies or countries, reciprocal visits to AI research facilities (with appropriate safeguards to ensure intellectual property protection), and prenotification of major model releases or significant capability upgrades.

Military AI governance may require separate protocols with different verification standards than civilian systems, though just how different is hard to say, and it is easy to imagine ways in which these standards will converge (e.g., if militaries are heavily dependent on privately produced AI capabilities and private com-

[11] Sastry et al., 2024.

[12] Raffaele Huang and Liza Lin, "Chinese AI Companies Dodge U.S. Chip Curbs by Flying Suitcases of Hard Drives Abroad," *Wall Street Journal*, June 12, 2025.

[13] Andrew J. Coe and Jane Vaynman, "Why Arms Control Is So Rare," *American Political Science Review*, Vol. 114, No. 2, May 2020.

[14] Michael C. Horowitz and Paul Scharre, *AI and International Stability: Risks and Confidence-Building Measures*, Center for a New American Security, January 2021; Sarah Shoker, Andrew Reddie, Sarah Barrington, Ruby Booth, Miles Brundage, Husanjot Chahal, Michael Depp, Bill Drexel, Ritwik Gupta, Marina Favaro, et al., "Confidence-Building Measures for Artificial Intelligence: Workshop Proceedings," arXiv, arXiv:2308.00862, August 3, 2023.

panies increasingly adopt rigorous security protocols inspired by military analogues).[15] Unfortunately, it does not seem likely that, on their own, externally visible signals will help to determine whether computing power is being used for civilian or military purposes or for authorized or unauthorized purposes, though there is some reason for optimism that a combination of software-based and hardware-based approaches could achieve versions of this.[16] For example, special-purpose hardware could securely and privately attest to the fact that certain properties of an AI system remain similar over time, or that certain guarantees are not violated, and this could be combined with more-direct analysis of the software in question.[17] By analogy, in-person inspection of a nuclear facility can only go so far, but it can be combined with other techniques, like portal monitors and satellite imaging, to reduce the likelihood of unnoticed, significant changes. Another possibility is that *compute accounting* could be used as part of a verification regime that ensures that large swaths of computing power that a company or country has access to is "spoken for" (i.e., how it was used is well known), and the amount of "dark compute" (that is not spoken for) can be reduced to acceptable levels.[18] These scenarios are made possible by the fact that computing hardware can carry out only a finite number of computations per second, and again, assuming a capital-intensive and concentrated supply chain, it is also feasible to know where the most-advanced chips reside.

It would be infeasible to ensure compliance for the long tail of AI systems and applications, but this is possible for frontier systems. Notably, it took years to negotiate, do fundamental research for, and execute historical arms control agreements where verification was a key component, so getting started sooner rather than later is ideal in the case of AI—as is finding ways to leverage AI toward verifying agreements, so that rapid technological progress can work with us rather than against us.[19]

Incentivizing Compliance

In cases where the compliance of other parties is not the primary reason for a party not to comply, several options are available to induce compliance. These include gating access to computing power (e.g., via fine-grained export controls with strict verification protocols attached, up to and including remote off-switches in the event of confirmed noncompliance or withdrawal from the auditing and verification regime discussed above); other trade-related actions, such as tariffs or embargoes;

It does not seem likely that, on their own, externally visible signals will help to determine whether computing power is being used for civilian or military purposes or for authorized or unauthorized purposes.

[15] Jürgen Altmann, "Verification Is Possible: Checking Compliance with an Autonomous Weapon Ban," *Lawfare* blog, April 8, 2024.

[16] Ben Harack, Robert F. Trager, Anka Reuel, David Manheim, Miles Brundage, Onni Aarne, Aaron Scher, Yanliang Pan, Jenny Xiao, Kristy Loke, et al., "Verification for International AI Governance," AI Governance Initiative, Oxford Martin School, University of Oxford, July 3, 2025.

[17] For example, see flexHEG, homepage, undated.

[18] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al., "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims," arXiv, arXiv:2004.07213, April 20, 2020; Yonadav Shavit, "What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring," arXiv, arXiv: 2303:11341, May 30, 2023.

[19] Mauricio Baker, "Nuclear Arms Control Verification and Lessons for AI Treaties," arXiv, arXiv:2304.04123, April 8, 2023; Aaron Scher and Lisa Thiergart, *Mechanisms to Verify International Agreements About AI Development*, Machine Intelligence Research Institute Technical Governance Team, November 27, 2024.

and the threat of military action. The credibility of rules being enforced in practice benefits from a gradation of possible responses, with clearly specified triggers for each.

Access to next-generation chips, chip design tools, or even large-scale cloud computing resources could be made conditional on adherence to agreed-upon safety and security standards, verified through the auditing regime discussed above. In cases of confirmed noncompliance or withdrawal from the verification regime, consequences could include ending exports to a certain company or country or remotely deactivating such hardware.[20] Careful design and innovation are needed to ensure that such an approach would not have unacceptable privacy and security properties.[21]

Other trade and technology-related actions can create significant disincentives for noncompliance. These could include targeted tariffs on AI-related or non–AI-related goods and services, restrictions on international financial transactions for noncompliant entities, visa revocations, or full trade embargoes. Implemented multilaterally for greatest effect, such sanctions would signal strong international condemnation and impose tangible economic costs on entities or states that flout agreed norms. By broadening economic pressure beyond the direct inputs to AI development, such measures could have a greater chance of changing certain countries' behavior.

In extreme cases, where noncompliance results in the development or deployment of AI systems that pose clear and imminent threats to international security (e.g., facilitating proliferation of weapons of mass destruction [WMDs], eliminating human oversight of AI systems conducting high-risk R&D), more-direct security-related incentives may come into play. This involves the credible threat of security actions by a coalition of compliant states.

There may appear to be a fine line between the preemptive strike scenario and the use of security threats to induce compliance, but there is a critical distinction: The latter makes reference to well-defined safety and security standards that have been violated, whereas the former emerges in part due to the absence of such standards.

A balanced approach that incorporates positive incentives, such as preferential access to certain AI capability insights and early access to commercial fruits of AI, should also be pursued where feasible. Ultimately, the goal is to create an environment where the real and perceived strategic benefits of participating in a well-governed, safety- and security-conscious AI ecosystem decisively outweigh any real and perceived advantages of unilateral, risky behavior. Preparedness, however, must also account for the possibility that some actors will remain outside such a regime, requiring continuous monitoring for AI misuse "in the wild" and investments in societal resilience.

Conclusion

An international crisis (or worse) is likely the default outcome from current approaches to AI development and deployment. Increasingly high-stakes development and deployment decisions in the private sector are yielding material changes in the competitive balance on a timescale of weeks and months rather than years, and the pace of change will only accelerate from here as faster-than-human AI contributes materially to the pace of innovation.

[20] Gabriel Kulp, Daniel Gonzales, Everett Smith, Lennart Heim, Prateek Puri, Michael J. D. Vermeer, and Zev Winkelman, "Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090," RAND Corporation, WR-A3056-1, 2024.

[21] James Petrie, Onni Aarne, Nora Ammann, and David Dalrymple, *Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees*, Institute for AI Policy and Strategy and Advanced Research and Invention Agency, August 23, 2024.

The current geopolitical landscape, marked by an acrimonious U.S.-China relationship and limited appetite for multilateral risk reduction, further exacerbates these risks. The prospect of an AI race playing out against the backdrop of potential flash points, like Taiwan, underscores the fragility of international security in this new technological era.

Action should be taken quickly to address the fundamental challenge at the intersection of AI and geopolitics: transforming unbridled competition into competitive coexistence. This requires a rapid, concerted global effort to define baseline safety and security standards, establish robust auditing and verification mechanisms for these standards, and create powerful incentives for compliance. The alternative is a path of increasing volatility, growing mutual distrust, and a high likelihood of disaster.

CHAPTER 4

One Does Not Simply Dismiss the Nuclear Revolution

James D. Fearon

Policy discussions of the potential impact of AGI on U.S. national security routinely begin with the suggestion that AGI will or could have impacts comparable to the nuclear revolution.[1] In this paper, I argue that this is not likely, at least not in the sense that AGI will confer military advantages that undo core effects of the nuclear revolution. In particular, I do not think that AGI will significantly increase the ability of an AGI-enabled state to take large amounts of territory from a nuclear-armed state.

The strategic essence of the nuclear revolution—mutual vulnerability to obliteration of major cities—is extremely hard to reliably escape, even in an AGI-enabled world. AGI-facilitated efforts to do so will foster concerted efforts to counter these, and there are many, many options, even for a state that does not have AGI. A state that does not have AGI but does have the ability to destroy some of an attacker's cities is dangerous to invade—even if the attacker has swarms of AGI-enabled drones or any other highly effective battlefield enhancements. It also has options to resist coercion and extortion on truly vital interests. AGI may create yet more ways to impose costs on the population of another country without having to defeat its army, but nuclear weapons can already provide this capability to a nearly world-ending degree (and, I am arguing, this is not likely to be undone by AGI).[2]

Much of the current discussion about policy responses to impending AGI focuses on "What if one state got AGI first and could use it to dominate us/others?" and "Given that, should we AGI-arms race like mad, or consider military preemption options, or maybe try to construct some kind of AGI arms control regime?" But if AGI is unlikely to end vulnerability to nuclear attack—whether via missiles, smuggling and pre-positioning in cities, or multiple other means of delivery—then it is hard to see how getting to AGI first can imply a discontinuous leap in classical military advantage that warrants present-day panic and highly costly or dangerous policy responses.

This is *not* to say that AGI would have no significant consequences for militaries, for national security, and for societies and economies. It certainly would. For one thing, if AGI has the effect of weakening secure

[1] This claim is made routinely about AI, let alone AGI. For example, "AI's integration across the full spectrum of military operations promises to revolutionize warfare even more fundamentally than nuclear weapons" (Bill Drexel, *Promethean Rivalry: The World-Altering Stakes of Sino-American Competition*, Center for a New American Security, April 2025, p. 1). Similar claims are too numerous to cite, but regarding AGI, see also Leopold Aschenbrenner, "Situational Awareness: The Decade Ahead," Situational-Awareness.ai, June 2024; and Dan Hendrycks, Eric Schmidt, and Alexandr Wang, "Superintelligence Strategy: Expert Version," arXiv, arXiv:2503.05628, April 14, 2025b.

[2] I am in general agreement with Edward Geist, whose excellent book provides a detailed analysis of the specific question of whether AI advances would be likely to make a "splendid first (nuclear) strike" possible (Edward Geist, *Deterrence Under Uncertainty: Artificial Intelligence and Nuclear Warfare*, Oxford University Press, 2023). He thinks not, for reasons that I treat below under headings of complexity and information requirements.

second strike under current force postures, the qualitative and quantitative arms racing that states undertake to restore assured destruction may come along with dangers and heightened risks, in addition to just being costly.[3]

But my guess would be that the most concerning national security impacts of AGI will take a different form if AGI happens. Namely, AGI would accelerate and worsen an already evident trend—specifically, the democratization of WMDs and mass subversion. Major risks and dangers will be associated with terrorism by individuals, non-state groups, and groups ambiguously connected with states. Biological and chemical weapons have been at increasing risk for terrorist use even before AGI. This would accelerate and worsen, as would the risk of lab accidents. AGI might also begin to democratize access to nuclear, or at least radiological, capabilities, increasing the likelihood of proliferation to both state and non-state actors.

On mass subversion, AGI would likely accelerate the ability of individuals and small groups to design and implement more-effective influence campaigns that proliferate distrust, undermining the ability of government and society to undertake net beneficial collective action, including the ability to deal with harmful or beneficial effects of AI and AGI. This is already happening due to information technology and changes in media. It does not even need that much help from AI.

Jim Mitre and Joel Predd identify five "hard national security problems" posed by AGI:

1. "enable a significant first-mover advantage via the sudden emergence of decisive wonder weapons"
2. "cause a systemic shift that alters the balance of global power"
3. "empower nonexperts to develop weapons of mass destruction"
4. "cause the emergence of artificial entities with their own agency to threaten global security"
5. "increase strategic instability."[4]

In this paper, I argue that (1) and (2) are not likely enough to warrant intensive arms racing or preemptive attacks. Problems (3) and to some extent (5) deserve more policy focus and efforts. I do not have considered views on (4).

Regarding (3), the threat goes beyond classical weapons and instruments of terror that cause physical harm, to new and improved means of undermining state capabilities and social capital. On the plus side, the United States and China could have a common interest in cooperating to limit proliferation of AGI-enabled terror capabilities. On the minus side, the upside dual-use potentials of AGI will make this extremely difficult to accomplish. Shadow propaganda "wars" between all sorts of actors, including states, will also be very difficult to mitigate by interstate cooperation.

Terms and Some Assumptions

In what follows, I am going to use *AGI* in a deliberately expansive manner, to cover a broad range of ideas that experts may distinguish and fight over. I will use it to refer to current or future capabilities of computers, their networks, and software that solve problems or execute tasks that until recently could only be done by humans or that humans are not capable of.

Others may want to distinguish between the following:

[3] For a broader discussion of this class of risks, see Paul Scharre, "Debunking the AI Arms Race Theory," *Texas National Security Review*, Vol. 4, No. 3, Summer 2021.

[4] Jim Mitre and Joel B. Predd, *Artificial General Intelligence's Five Hard National Security Problems*, RAND Corporation, PE-A3691-4, February 2025, p. 2.

1. AI, in the sense of machine learning methods and large language models, including in the form of the frontier models that currently exist
2. AGI, in the sense of breakthroughs in these areas that create human-level intelligence capabilities (something like more generalized problem-solving or analytical capability)
3. superintelligence, in the sense of breakthroughs that create AGI that vastly exceeds human analytical capabilities.

For my purposes, since claims about revolutionary political and military implications can be and have been based on any of these, I would like the arguments that follow to be responsive to the range from (1) through (3). For convenience, I will use *AGI* even when referring to hypothetical capabilities that would be seen as exhibiting "superintelligence."

The arguments that follow do not depend critically on how soon (2) or (3) arrives, although it is certainly plausible that shorter timelines would be more destabilizing by creating greater risks of panicked responses and harmful applications—especially, I think, from "escape," whereby unregulated access allows innovation of malicious uses.

"Superintelligence" merits more comment. How can one constructively define something that, by hypothesis, we cannot currently conceive of? I do not think we can, which makes me doubtful of the value of trying. "Vastly exceeds human capabilities" is more of a placeholder than a definition. For thinking through policy implications, it may be more productive to begin by asking what superintelligence, whatever it might end up being, could *not* do (or would likely struggle to do). The next section offers some ideas here.

AGI Calvinball?

If one is allowed to hypothesize literally any capability for AGI consistent with the laws of physics, then it becomes hard to have a meaningful discussion about courses of action for policy. There are then infinite possible dire threats we can imagine and no way to protect against them. From this starting premise, the only rational course of action would be to try to immediately destroy any and all work in this area, by force if necessary, everywhere. It would not make much sense to race to try to acquire these God-like powers first, because no accountability mechanisms could be secure enough to trust even one's own government or leaders with them.[5]

To illustrate more concretely, here are some scenarios in the nuclear domain:

- **AGI threat 1:** What if AGI could quickly figure out how to electronically infiltrate and take control of another country's nuclear command, control, and communications (NC3) system, to disable, destroy, or commandeer it?
- **Response 1:** Thoroughly and rigorously air gap these systems.
- **AGI threat 2:** What if AGI can by bypass air-gapping by quickly figuring out how to identify, brainwash, or extort relevant individuals in an NC3 system? Indeed, what if AGI can figure out how to brainwash or extort a country's leaders?
- **Response 2:** Okay, not sure what could be done in that case. (But also not sure nuclear is even a top-five problem in this event.)

[5] Davidson, Finnveden, and Hadshar discuss the risk of "AI-enabled coups" where "leaders of frontier AI projects, heads of state, [or] military officials" create an AI workforce that is "singularly loyal," has hard-to-detect loyalties to one person, and may control robot and drone armies (Tom Davidson, Lukas Finnveden, and Rose Hadshar, *AI-Enabled Coups: How a Small Group Could Use AI to Seize Power*, Forethought Foundation, April 2025, p. 4).

- **AGI threat 3:** What if AGI quickly solves the problem of perfect missile defense, enabling the Star Wars shield over a large country that President Reagan envisioned? Then secure second strike no longer exists, and the nuclear revolution is canceled.
- **Response 3:** I am not sure if this is truly compatible with laws of physics and engineering realities.[6] But even supposing major progress in the direction of effective national missile defense, nuclear weapons do not have to be delivered by missiles. For example, faced with greatly improved missile defenses, nuclear-weapon states could then have incentives to infiltrate and pre-position small nuclear devices within the AGI-enabled adversary. If huge volumes of drugs can be smuggled into a country despite great efforts against this, why not small nuclear devices? Why not take advantage of drugs or other contraband smuggling for just this purpose?
- **AGI threat 4:** AGI will be able to detect and prevent any such smuggling and pre-positioning attempts because its pattern recognition, intelligence, and data analysis capabilities will be so great.
- **Response 4:** That would really be something.

It is possible to put *some* bounds on hypothetical AGI capabilities, and this may be worth speculating about even if the bounds are not at all tight (nor can they be):[7]

- **Laws of physics**, as noted. For example, AGI will not invent faster-than-light travel. We have too many good reasons to believe that this is not possible.
- **Randomization.** Suppose a general chooses among multiple attack vectors by a physical randomizing device (like coin flips) in a private space. AGI cannot anticipate the direction or mode of attack at the very least until the order goes out. Prediction of coin-flip results is hypothetically possible and consistent with the laws of physics (by classical mechanics) if the AGI had enough detailed information, but that information is not available to it.[8]
- **Complexity and information limits.** Generalizing the last class of cases, AGI will not be able to make point predictions of future events that result from complex processes. For example, will there be a war between the United States and China in 2028, or would a war escalate to nuclear use if it happens? Even if one believes in Laplacian determinism, the information requirements are too vast and outcomes are too sensitive to complex, effectively invisible interactions whose numbers increase exponentially by the day as we try to look further into the future.[9]

  This class of cases is relevant for the question of whether AGI could render a successful preemptive strike on an adversary's nuclear weapons. If the adversary has postured these for a secure second strike, an enormous number of things must go just right for preemption to work for sure or with high

[6] See, for example, hypersonics and the inability to test at scale an extremely complex system with thousands of hardware and software and probably human fail points. Geist (2023) develops essentially this argument based on the problems of sensor fusion in particular. More below.

[7] Edward Geist and Alvin Moon, "What Even Superintelligent Computers Can't Do: A Preliminary Framework for Identifying Fundamental Limits Constraining Artificial General Intelligence," RAND Corporation, WR-A3990-1, 2025. Geist and Moon discuss constraints related to computational complexity and laws of physics.

[8] What if AGI can suborn all of your generals or turn them into Manchurian-candidate generals without their awareness? This capability would imply problems that go way beyond use on military leaders.

[9] James D. Fearon, "Causes and Counterfactuals in Social Science," in Philip E. Tetlock and Aaron Belkin, eds., *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*, Princeton University Press, 1996. As an analogy, ask yourself if AGI will be able to predict the winning number in next week's Powerball lottery or whether it will be raining in Beijing on May 5, 2026.

probability. The situation is "one off," idiosyncratic, and untestable—unlike, say, running a complex national air traffic system.

What about probabilistic forecasts? In the military realm, we are generally interested in predictions about events where there can be no real test data on which to base probability estimates. The best one can do is models and simulations, which are necessarily built on myriad assumptions and simplifications due to complexity and lack of information. These can be useful, and AGI might be able to make them better. But the nature of problems means that it will be very difficult to tell if the forecasts and predictions are any better or are in fact highly misleading.[10]

- **Social engineering and complex public policy problems?** We already know that AI in the form of machine-learning methods can solve or make great advances on well-defined and thus relatively narrow computational problems, like chess, image recognition, and protein-folding. This will continue, bringing enormous potential benefits along with enormous dangers in biotechnology and automation of production processes (for example). In some areas, "solve computation/optimization problem X" can lead almost immediately to real-world effects, such as a new drug or vaccine.

  But in areas where desired outcomes require re-coordinating the actions of large numbers of people, a different set of challenges appears. These include complexity and information limits, as above, but now add in problems getting large numbers of people and organizations to re-coordinate their daily activities and choices. Suppose that AGI has brilliant ideas about, for example, a new weapon system—say, how to make drone swarms work incredibly well. To become real, the ideas still have to be implemented in the form of factories, production processes, and a thousand tiny policy decisions that go along with these. Once you have the new tech, implementing it as an operational capability in a military entails another layer of thousands of decisions and choices.

  These myriad small design and process decisions and choices inevitably come along with conflicting preferences among people and different parts of organizations, as well as tons of private information and intrinsic uncertainties about preferences and costs and benefits regarding the many decisions and choices. Just as we know that faster-than-light travel is not possible, we know that there are hard constraints on collective choice processes for aggregating decentralized information into collective action.[11]

  A related type of constraint may be less relevant in the national security sphere but should be noted. "Solving" most public policy problems requires normative standards for comparing states of affairs, and what the right standards should be cannot be "solvable" by AGI. For example, "Design an optimal national health care system." For a policy area with large distributional elements, AGI cannot solve the question of what is the normatively best distribution.

Nuclear Revolution Refresher

Eighty years later, the nature and implications of the nuclear revolution remain poorly understood, or neglected, even in debates within defense and foreign policy expert communities.

[10] Geist (2023) makes a stronger claim about a specific military challenge, arguing that results in theoretical computer science imply that we can be certain that AGI will not be able to solve the sensor fusion problems required to make a "splendid" nuclear first strike possible. The reason is basically that suggested above: Lack of information plus complexity make some problems computationally intractable, no matter how good a computer you have (consistent with laws of physics).

[11] For example, AGI cannot identify a way to aggregate individual preferences that satisfy Arrow's weak desiderata for collective choice, because this is not possible by Arrow's theorem. Nor can it identify a mechanism that guarantees truthful revelation of preferences for any and all social choice problems (Gibbard-Satterthwaite theorem, approximately).

One of the most common errors is to associate the nuclear revolution with the huge increase in the destructive power that thermonuclear bombs supply. Yes, but as Thomas Schelling argued in "The Diplomacy of Violence," the revolutionary implications of this power really come from the fact that the weapons can be delivered on an adversary's population without needing to defeat its military first.[12] Further, there is no reliable defense and no reliable means of preemption against a state that has invested in secure second-strike capabilities.

What follows? North Korea's population is less than half of South Korea's, 7 percent of the United States', and 2 percent of China's. Its economy is minuscule compared with any of these. But its nuclear forces render it quite secure against invasion.[13] In the pre-nuclear world, a state like North Korea would be basically helpless against a determined major power, unless it had a committed major power ally. By contrast, nuclear North Korea does not need a major power ally to deter invasion and forcible overthrow of the Kim regime.

In the pre-nuclear world, the term *balance of power* meant something. This is because in the conventional world, a state's ability to defend itself against invasion depended on the size of its armed forces—numbers of soldiers, tanks, aircraft, and so on—*relative to* an adversary's forces. In the nuclear world, the fact that the United States has thousands more nuclear warheads (and tanks and soldiers) than North Korea does not make it any more possible for the United States to invade and depose the regime, provided that North Korea's weapons are postured so as to ensure a reasonable second-strike capability. Nor has the United States' huge advantage in numbers of nuclear weapons and other forces given it any noticeable ability to coerce the Kim regime to stop its nuclear weapon development or any of its other malign foreign policy behaviors.

It also does not matter if the United States or China or South Korea, or all three of them, has a vastly larger economy than North Korea. In the pre-nuclear world, a state's ability to defend itself from invasion and coercion depended crucially on its economic size (and alliances), since relative forces mattered and force size and quality depended ultimately on having a large population and an advanced economy. No more, for countries that can acquire minimally adequate nuclear capability. This means that even if AGI allows a country to create massive economic growth, this does not give it a meaningful military advantage over a nuclear-armed adversary, at least not in the classical military sense of an ability to convert economic might into successful military operations to take territory.

To be clear, the nuclear revolution does *not* eliminate the possibility of major armed conflict, even between nuclear states with secure second-strike forces. The nuclear revolution in effect changes the strategic form of intense conflicts between major powers from *wrestling matches*, whose outcomes depend on relative size, strength, and skill, to *auctions*, whose outcomes depend on which side is willing to run a higher risk of nuclear escalation or to incur piecemeal nuclear damage.

In an intense crisis or nascent war between nuclear powers over vital interests, the states have no choice but to effectively "bid" in nuclear risk or limited nuclear use, since they retain this capability even after suffering losses on the conventional battlefield. In a wrestling match, the stronger party may be able to throw the other even if it cares less about what is at issue. In a nuclear auction, the ability of the conventionally weaker (or losing) side to raise the ante by nuclear risk or damage means that the balance of interests becomes at least as important as the conventional balance of power.[14] While rolling the tanks into the adversary's capital becomes very unlikely, the risk of a nuclear war cannot be eliminated and becomes more likely the more intensely the states care about whatever is at issue (e.g., Berlin, Cuba, Kashmir, Taiwan).

---

[12] Thomas C. Schelling, *Arms and Influence*, Yale University Press, 1966, Ch. 1.

[13] Indeed, Seoul's proximity to the North renders it dangerously vulnerable to a conventional missile barrage.

[14] For instance, Kim Jong Un plausibly cares more about keeping power, or would be willing to launch missiles if he thought he were losing it, than China, the United States, or South Korea cares about ending his family's rule.

The Nuclear Revolution Is Not to Be Trifled With

To repeal the nuclear revolution and its main effects, AGI would need to allow its master to either (a) undertake a successful, disarming preemptive strike on a nuclear adversary or (b) develop a perfect or close-to-perfect defense against nuclear attack. Neither seems likely unless we are allowed to imagine that AGI enables highly effective targeted mind control and that one side attains and can implement this capability before the other.

Splendid first strikes? Before cyber, the only way to execute a disarming first strike against an adversary's deployed nuclear weapons would have been kinetic attacks. This was and remains a massively complex problem against a country with a well-developed nuclear triad or even a highly secure subset of the triad. An enormous number of things have to go just right to completely eliminate a real possibility of nuclear retaliation.

Consider the challenge of finding and successfully destroying multiple submarines at the same time, and likewise for road-mobile platforms, bombers (which can be hidden or kept airborne), and hidden or deeply buried caches of weapons that can be loaded onto them. Even with phenomenal sensor fusion and analysis capabilities enabled by AGI,[15] the effort would necessarily depend on a massive of amount of hardware functioning just as intended, from the sensors to the missiles to the people needed to undertake any nonautomated task—of which there would be thousands—or, if these were automated, then massively more hardware and software subject to failures. While some degree of minor equipment failures might be a tolerable risk, I cannot see how AGI could reliably estimate a probability of success for an event that requires the concatenation of so many thousands of uncertain steps and for which no real-world testing is feasible. As Geist argues, the problem can be made arbitrarily more difficult by proliferating decoys and other (probably AI-enabled) means of deception.[16]

Further, if sufficiently concerned about an adversary's first-strike capabilities from missiles, a technologically sophisticated state could pre-position and hide nuclear weapons in an adversary's country, in the oceans, or in remote space orbits. This is not purely hypothetical. Even now, before AGI, Russian concerns about U.S. missile defense and U.S. withdrawal from the Anti-Ballistic Missile Treaty appear to have contributed to Russian efforts to develop means of nuclear delivery not susceptible to missile defense. These include Poseidon, a nuclear-powered, very fast, autonomous nuclear torpedo capable of traveling long distances, which could be carried on submarines or potentially hidden in pre-positioned seabed locations.[17]

Since cyber, there may be new ways that a superintelligent AI system could disarm a nuclear competitor without classical kinetic attacks—namely, by destroying or disabling command, control, communications, and computers (C4) or the weapon platforms themselves (e.g., Stuxnet). The first way (destroying or disabling C4) would be temporary, so to truly remove the target's nuclear capabilities, this would have to be accompanied by some kind of kinetic attack or government overthrow. The second would also likely be temporary, although in principle the time to regenerate destroyed forces could be long.

Nuclear weapon systems must already be air-gapped to a significant degree. Increased threat of AGI/cyberattack would make it sensible to carry out air-gapping much more rigorously and carefully. Large parts

[15] Geist (2023) argues that the problem of sensor fusion in this problem set may be computationally intractable, even with quantum computers.

[16] Geist, 2023.

[17] See Anya L. Fink, *Russia's Nuclear Weapons: Doctrine, Forces, and Modernization*, Congressional Research Service, R45861, April 21, 2022. On the seabed idea and suggestions that Russia may be pursuing it, see Felix Lemmer, "Poseidon: Oceanic Multipurpose System Status-6, Kanyon," Hertie School Centre for International Security, March 2022. Andrew Lim (personal communication) observes that if one is worried that AGI would lead to "transparent oceans," the United States, Russia, and China could position ballistic missile submarines in Lake Superior, Lake Baikal, or Lake Qinghai, thus beyond the reach of antisubmarine warfare.

of the C4 systems must of course depend on electronic communications, and some parts must depend on encrypted radio-frequency transmissions.[18] I would be surprised if there were not clever ways to make those pathways impenetrable—for example, by using high-level encryption combined with hard-copy randomized tables of codes known to senders and receivers.

In sum, it is not likely that AGI could produce high confidence that a disarming first strike could succeed against a state with well-postured nuclear forces.

Even if the nuclear revolution remains robust in this sense, one might still ask about the implications of marginal improvements in first-strike capability. Hold fixed state A's nuclear posture and C4ISRT (command, control, communications, computers, intelligence, surveillance, reconnaissance, and targeting) systems. It is certainly possible that the introduction of AGI capabilities could make state B's nuclear first-strike option against A *more* likely to succeed than was previously the case, in the sense of reducing expected nuclear damage from retaliation. Nuclear history suggests that this will encourage state A to take countermeasures (which can be effective, per above), with its speed and effectiveness depending on many political and organizational circumstances. As noted in the introduction to this paper, if AGI has the effect of weakening secure second strike under current force postures, the qualitative and quantitative measures that states undertake to restore assured destruction may come along with dangers and heightened risks, in addition to just being costly.

How likely is it that a leader's belief in improved nuclear first-strike prospects would, in the short run, occasion a premeditated attempt? My guess would be not likely, both for the complexity and untestability reasons already discussed, and because I do not think there are many circumstances in which a leadership would see political and military advantage from a first-strike attempt as worth the risks and costs. More likely would be risk coming from reciprocal fears in an intense crisis or nascent war, a type of risk that can be increased by first-strike temptations. Neither can be completely ruled out, of course, but what to do about it except work to make second-strike capabilities more robust (which is possible)? How much to fear and plan against crazy or highly risk-acceptant nuclear-armed leaders has been a question of intense debate since the nuclear age began.

Perfect defense? The physics and engineering obstacles to perfect missile defense are even worse than those for designing and executing a "splendid" nuclear first strike. And in this area, opponents have many excellent options for designing around or overwhelming any given missile defense system, even if these were to improve greatly with AGI.

For example, nuclear bombs can be delivered by hypersonic missiles or a fractional orbital bombardment system (FOBS) whose trajectories can change unpredictably (randomization again). Increasing the number of hypersonics or FOBS devices creates a sensor fusion problem that almost surely becomes computationally impossible very quickly. Using decoys lowers the costs of these options and can massively increase the problem faced by a defender.

It is not likely that AGI could produce high confidence that a disarming first strike could succeed against a state with well-postured nuclear forces.

Worried that AGI could figure out how to commandeer an attacker's communications with the missiles (e.g., targeting info, guidance)? In a pinch, they can be programmed in air-gapped settings before launch and not require subsequent contact, dangerous though this would be (think Dr. Strangelove).

[18] I do not know how important this is. But for submarines and road-mobile launchers in some terrains, acoustic or radio-frequency signals would be a necessity.

As indicated earlier in this paper, even if we granted AGI the power to bring about a perfect Golden Dome against missile attacks, there are other ways to deliver nuclear bombs against an adversary's population. Near-perfect missile defense would strongly incentivize other approaches, such as smuggling and pre-placement of miniaturized devices or uncrewed undersea vehicles to be released from submarines or commercial vessels near important ports. We can postulate AGI capabilities to try to predict and forestall such threats, but I do not see how capabilities short of highly effective and near-universal mind control could really guarantee success or allow meaningful estimation of the probability of success against a highly motivated adversary.[19]

Conventional Considerations

Even if AGI does not end mutual assured destruction among nuclear powers—and the revolutionary implications that follow from this—AGI would of course have many classical military applications. By *classical*, I mean applications that render weapon systems and combined arms warfare more effective at taking or defending territory against a non-nuclear opponent or nonvital, marginal territory against a nuclear opponent (in some cases, depending on the balance of interests). Just as being an early adopter and exploiter of drone technology helped Azerbaijan conquer Nagorno-Karabakh and Ukraine stymie Russian advances, first exploiters of AGI-driven battlefield enhancements may be able to seize or prevent seizure of marginal territory valued for nationalist reasons. This is not nothing, of course, but it would be nothing particularly new in the history of military technology compared with the nuclear revolution.

Two AGI applications particularly worth speculating about are, first, greatly improved targeting capabilities for conventional munitions (scale, speed, and accuracy), and second, cyber penetration of all manner of systems needed for military operations.

If obtained first by a revisionist state, AGI-enhanced targeting could facilitate seizure of territory in some contexts, such as a People's Republic of China (PRC) attempt to disable U.S. capabilities in the early stages of an attack on Taiwan. In fact, vulnerability of U.S. assets in the Western Pacific to missile attacks is already a big problem given PRC missile capabilities and U.S. force posture. Even with current technology, however, U.S. and partner missile capabilities can be rendered much more secure by following the same principles behind ensuring secure second strike as in the nuclear domain.[20]

AGI successes could make the challenge of secure conventional (missile) second strike more difficult. There will be responses, however, for the same reasons as argued for the nuclear case. Submarines, road-mobile launchers, and bombers can be postured in ways that make targeting them a computationally extremely difficult problem. This is an area where time lags and speed of implementation could matter; it would help if the defender has time to build and adjust or to take advantage of the same AGI capabilities to counter more effectively.

Already happening with existing software, the second application that AGI would seem likely to accelerate is cyber penetration of an adversary's military systems, perhaps especially C4ISRT. An interesting implication of this mode of warfare is that it should tend to weaken confidence in one's ability to execute almost any complex military operation, whether on defense or offense. One does not know if one's systems are consequentially penetrated, or if, when something does not work right, this is because of an adversary exploit or

[19] The recent Ukrainian drone attack on Russian airfields involved smuggling the weapons deep into Russian territory (by the way).

[20] Andrew S. Lim and James D. Fearon, "The Conventional Balance of Terror: America Needs a New Triad to Restore Its Eroding Deterrence," *Foreign Affairs*, April 22, 2025.

a fault of your own. If roughly symmetrical, this could actually be a mild force for peace by making initiating operations more risky.

But an *asymmetrical* AGI capability to ferret out any and all adversary cyber exploits—and at the same time disable adversary C4ISRT—could greatly empower conventional operations, at least in the short run. I do not know if this is a realistic prospect or whether it could be done without the adversary knowing in advance (which would be relevant for surprise and thus impact). Even with current technology, investing in redundant communications and secure control systems should be a priority.[21]

What will the impact on the offense-defense balance be for conventional operations, if and when AGI becomes available to more than one state? Even putting aside the "gumming up the works" cyber effect just mentioned, my guess is that defense would be favored, because AGI would effectively increase firepower and targeting capabilities—as already suggested by effects of unmanned aerial vehicles, which AGI could make more potent still. For taking territory, attackers have to cross ground, exposing themselves in the process. Why would improved battlefield surveillance and targeting favor suppressive fire relative to fires directed at attacker units?[22]

In practice, there will almost always be time lags between AGI ideas and concepts for operations, on the one hand, and implementation on the other. This is because physics, engineering, production processes, and human or robot adaptation or exercises intervene between idea and use. An implication is that the most important military effects of AGI may lie in who figures out how to use it to make their *defense industrial base* radically more efficient and effective than anyone else's. That would mean solving complex, distributive political and organizational problems, which probably have more to do with societal and governmental capability than AGI (see above on constraints). Further, as argued above, the nuclear revolution means that having an amazing defense industrial base does not give you any great ability to take over, or even coerce, a North Korea, for example.

Terrorism, Trust, and Subversion

It is obvious that if AGI afforded individuals everywhere the ability to design, build, and release deadly and highly infectious pathogens, then more than one of 8 billion humans would try to do this. We would have to hope that AGI would also be great at designing antidotes and vaccines, that this is an area where states have sufficient common interest to try to lower the likelihood through AGI governance (controls), and that controls are actually feasible given the nature of the technology and the powerful commercial imperatives driving its development.

I am slightly less concerned about states developing and using chemical, biological, or radiological weapons with AI or AGI help than I am about individuals, non-state groups, and lab failures. While states can have incentives to explore chemical and biological weapon capabilities, it is fortunate that actual use can be subject to mutual deterrence, international arms control, and operational obstacles, such as how to keep your own population from getting sick.[23] Perhaps AGI could innovate in designing highly targeted pathogens. Human

[21] Presumably, if AGI enables qualitative leaps in cyber offensive capabilities, it would also imply a qualitative leap in defensive capability, such as the ability to quickly identify and disable all adversary exploits. Then, with symmetrical AGI capabilities, the net impact is not clear.

[22] Also, what about much more sophisticated mines that can be readily controlled by a defender?

[23] On chemical weapons, see James D. Morrow, *Order Within Anarchy: The Laws of War as an International Institution*, Cambridge University Press, 2014; he emphasizes mutual deterrence for the case of World War II. Some states appear to have invested heavily in bioweapons research. As best I can tell, there is no consensus view on why there has not been more military use. Koblentz gives multiple reasons to expect increasing use of bioweapons by state and other actors, which thankfully

genetic variation and the complexity of immune systems suggest that this might be hard to make work well, but it is certainly a risk.[24]

AGI will not lead to repeal of the nuclear revolution and its major consequences for international politics and national security.

Individuals, non-state groups, and states would all be tempted to use AGI to render propaganda and influence operations more effectively. This is already happening simply with social media. To date, a principal effect in the more democratic countries has been to create dozens of solipsistic narratives about social and scientific reality, in turn amplifying conflicting preferences and mobilization around these in society, while at the same time increasing skepticism that anyone can be trusted.

If AGI would be able to quickly get huge majorities of people "on the same page," with that page being literally whatever it or its master wanted, this would be a God-like power against which nothing much could be done. I view this as fanciful, with the more likely outcomes being even more skepticism and division. Perhaps AGI would make campaigns of disinformation more virulent, exacerbating an already growing set of problems.

Conclusion

AGI will not lead to repeal of the nuclear revolution and its major consequences for international politics and national security. One of these consequences is that it is extremely difficult to use military force to invade and conquer a nuclear-weapon state if its forces are properly configured. This means that the amount of coercive leverage that acquisition of AGI would afford, at least against nuclear states, like the United States or China, would be limited. Coercive threats of any kind, on sufficiently important matters to the target, can be countered with nuclear risk or piecemeal nuclear use. AGI may invent more ways to remotely harm another country's population (adding to nuclear, cyber, and long-range conventional missiles), but this would simply add to the condition of mutual hostages that already exists.

If this analysis is on target, then there is less reason to take costly or extreme policy measures based on the fear that another state (China) getting AGI first would "enable a significant first-mover advantage via the sudden emergence of decisive wonder weapons" or "cause a systemic shift that alters the balance of global power."[25] Neither new weapons nor greater economic productivity translates through military action to territorial or government control.[26]

we have not seen so far (Gregory Koblentz, "Pathogens as Weapons: The International Security Implications of Biological Warfare," *International Security*, Vol. 28, No. 3, Winter 2003–2004; Gregory D. Koblentz, *Living Weapons: Biological Warfare and International Security*, Cornell University Press, 2009). An important obstacle for non-state actors may have been that even if one learns how to make a chemical or biological agent, manufacturing and delivering it at industrial scale involves a new set of difficult challenges and is more susceptible to monitoring and detection. A possible danger from AGI would be if it enables easy innovation of highly infectious agents that spread themselves, *12 Monkeys* or virus-gain-of-function style. See more generally Roger Brent, T. Greg McKelvey, Jr., and Jason Matheny, "The New Bioweapons: How Synthetic Biology Could Destabilize the World," *Foreign Affairs*, August 20, 2024.

[24] In a recent review of AI and biosecurity risks, Wheeler says that "experts remain divided on the feasibility of effective pandemic pathogen design" (Nicole E. Wheeler, "Responsible AI in Biotechnology: Balancing Discovery, Innovation and Biosecurity Risks," *Frontiers in Bioengineering and Biotechnology*, Vol. 13, February 4, 2025, p. 4).

[25] Mitre and Predd, 2025, p. 2.

[26] AGI might lead to developments in missile defense or coordinated targeting that would drive more nuclear arms racing and costly or dangerous changes in posture and control mechanisms.

In terms of physical security, I expect that the bigger impacts would be in furthering individual and small group access to WMD capabilities ("empower nonexperts to develop weapons of mass destruction," in Mitre and Predd's list[27]). Perhaps this class of threats is amenable to mitigation by cooperation on AGI governance and self-regulation, because states can have strong common interests here. Just as the United States and the Soviet Union came to realize that they had a common interest in nuclear nonproliferation, perhaps the United States, China, and other states that get to AGI frontiers may have a common interest in regulating and containing AGI-powered chemical and biological engineering capabilities. In sharp contrast to the nuclear example, however, the dual-use problem is far more severe in this area, since there are probably enormous benefits available from permitting wide scientific access and great difficulty monitoring for prohibited work. Along these lines, Volpe argues that biotechnology falls in "the dead zone" for verifiable international cooperation (on which, see Vaynman and Volpe).[28]

Even less amenable would be international cooperation to mitigate the increases in misinformation, propaganda, and subversion capabilities that AGI would accelerate. These capabilities tend to undermine the social and political requisites for translating AGI innovations into either good or bad outcomes, when these require large-scale collective action.

---

[27] Mitre and Predd, 2025, p. 2.

[28] Tristan A. Volpe, "Biotechnology and the Dead Zone for Managing Dual-Use Dilemmas," in Nathan A. Paxton, ed., *Disincentivizing Bioweapons*, Nuclear Threat Initiative, 2024; Jane Vaynman and Tristan A. Volpe, "Dual Use Deception: How Technology Shapes Cooperation in International Relations," *International Organization*, Vol. 77, No. 3, Summer 2023.

CHAPTER 5

Averting Attacks Against AGI Development: Three Strategic Approaches

Karl P. Mueller

Will the ongoing race to develop AGI or other advanced AI lead to international instability—that is, will it result in hostilities between countries that would not otherwise occur? For the AI race to lead to an upsurge in interstate conflict, two things would have to happen.[1] First, the competition or the AI resulting from it would need to create or exacerbate insecurity, hubris, perceived invincibility, hatred, confusion, or some other combination of motivations in national leaders that would make them interested (or more interested than they were before) in using force against other states. One can envision a number of ways this might occur if we assume that AGI (or its precursors) will lead to expectations of dramatic economic advantages and new or enhanced military capabilities for those that possess and employ it effectively.[2] Second, deterrence would need to fail: Prospective attackers would have to decide that using force was a good idea for addressing their problems or achieving their goals in light of their beliefs about its potential consequences and risks.[3] Because discussions about the stability implications of advanced AI often concentrate on the first dynamic (AI incentivizing aggression), this paper instead focuses on the latter one (the decision to act).

I define *AGI* as advanced AI that can perform a broad range of complex, important tasks better or more efficiently than humans and, crucially, is capable of the sort of autonomous self-improvement and application creation that leads AGI enthusiasts to project explosive economic growth and transformational military capabilities following soon after its development. I do not expect AGI to appear prior to the 2030s and assume that more than a few years will be required for it to be widely integrated into major economies, militaries, and national leader decisionmaking following its arrival.[4] The rate of progress toward AGI emergence seems likely to be difficult for outsiders to anticipate or monitor with precision. However, as the following discussion will suggest, when considering deterrence and stability implications, the rate at which we approach the AGI threshold and the implications of reaching it (or of an adversary reaching it first) matter less than decisionmakers' beliefs and expectations about those matters.

---

[1] This is setting aside the possibility of AI gaining autonomous control of military capabilities on a large scale and independently starting wars, which at least in the near term seems unlikely and avoidable.

[2] Zachary Burdette, Karl Mueller, Jim Mitre, and Lily Hoak, "Six Ways AI Could Cause the Next Big War, and Why It Probably Won't," *Bulletin of the Atomic Scientists*, July 15, 2025.

[3] This presumes that wars are not started by accident, a proposition for which the historical evidence is strong. See Geoffrey Blainey, *The Causes of War*, 3rd ed., Free Press, 1988; and Erik Lin-Greenberg, "Wars Are Not Accidents: Managing Risk in the Face of Escalation," *Foreign Affairs*, October 8, 2024.

[4] See Kahl and Mitre (2025).

The pathway to conflict most relevant to the AGI race would be preventive attack: a state seeking to halt or impede a rival's real or imagined progress toward more advanced AI.[5] This might be especially likely if a competitor appeared to be on track to "win" the race and thereby become far more powerful, perhaps to the point of achieving monopolistic control of the new technology. The incentives for such action might appear very strong to leaders who expected that falling behind (or too far behind) in the AI competition could threaten the survival of their state or its position as a leading power.[6] Preventive attack could also appeal to a state that had developed AGI and was determined to preserve the advantage it had gained over its competitors.[7]

As the global leader in the development of frontier AI models, the United States will presumably want to deter China and other competitors from taking violent or other extreme action to derail U.S. technology firms' progress toward AGI, both to avoid suffering harm (to those AI efforts and more generally) and to avoid an event that could escalate to a larger and more costly conflict. Likewise, China will want to deter the United States or anyone else from conducting preventive attacks against its more centrally controlled AI enterprise. Since there are many policy options that might help to deter preventive strikes, this paper offers a first-order classification framework to help think about how to construct a strategy from them.[8] It is important to acknowledge before proceeding, however, that China's leaders might not have any interest in taking preventive action to impede U.S. AI progress. For example, they might pursue a "fast follower" strategy of exploiting U.S. trailblazing in AI development, content to see the Americans win the race to AGI in the expectation that China will be better able to operationalize the technology. Under such a strategy, preventive action might appear counterproductive to Beijing, while espionage to collect the fruits of U.S. AI research would presumably be a priority. Chinese leaders might also doubt Washington's ability to effectively control and wield AGI developed by a U.S.-based AI lab.

The pathway to conflict most relevant to the AGI race would be preventive attack: a state seeking to halt or impede a rival's real or imagined progress toward more advanced AI.

[5] I focus on preventive attack here because it is the pathway that appears most closely linked to national decisions about AI policy and the key debates surrounding them. However, AI-related targets could come under attack in conflicts not primarily caused by concerns about AI. For example, in a war between China and the United States over the fate of Taiwan, either of the major powers might threaten or attack the other's AI sector for coercive purposes or to degrade its rival's military or economic power. The strategy discussion in this paper would apply to those cases as well, and indeed many of its general points are relevant to deterring attacks beyond AI.

[6] Mueller identifies four categories of hypothetical capabilities that AGI advocates have argued the technology could provide to a state possessing it (Karl P. Mueller, *Heeding the Risks of Geopolitical Instability in a Race to Artificial General Intelligence*, RAND Corporation, PE-A3691-12, July 2025, p. 2):

- "offensive military and cyber power against which other states will be unable to protect themselves"
- "strategic defenses that will provide invulnerability against enemy attack, particularly a large-scale nuclear strike"
- "explosive economic growth that will transform the international distribution of power in the state's favor"
- "tools for information control that can effectively manipulate and reprogram adversary political systems."

[7] Zachary Burdette and Hiwot Demelash, "The Risks of Preventive Attack in the Race for Advanced Artificial Intelligence," RAND Corporation, WR-A4005-1, 2025.

[8] Although I will focus on the United States deterring preventive attacks by China and vice versa, most of what follows also applies to deterring other potential state or non-state attackers (and even to the possibility of internecine attacks by AI companies against their commercial rivals).

Preventing Preventive Action

Preventive attacks are actions taken by a state to weaken an adversary in advance of an anticipated or potential conflict or to eliminate a threat altogether, using armed force or associated means, such as cyberattacks, rather than tools of normal peaceful competition.[9] Preventive attacks against AI development could be directed against AI hardware and supporting infrastructure—labs, chip foundries, data centers, electric power generation or transmission, or facilities involved in developing or manufacturing products incorporating AI—or against the models themselves. They might also physically or coercively target people involved in the AI ecosystem, ranging from AI researchers to investors, to disrupt their activities.[10] The potential means of preventive attack are similarly varied, including but not limited to conventional military strikes (particularly by long-range missiles), cyberattacks or manipulation, sabotage (a wide spectrum from explosively overt to deeply subtle), assassinations, coercive threats, and psychological and information warfare. Nearly all of these methods are familiar from the record of preventive actions taken or contemplated against states developing nuclear weapons or, as demonstrated by states including Russia, Ukraine, and Israel in their recent conflicts, adversaries' development or acquisition of potent conventional weapons.

This diversity of potential targets and attack methods leads to a correspondingly extensive range of possible defensive and deterrent measures.

Two other dimensions of variation among threats figure prominently in devising strategies to avert preventive attacks. One is the extremity of the preventive action; how destructive and how outrageous an act is will play a large role in shaping the range of potential responses to it, which will in turn affect an attacker's choices about which options to consider. The other is how readily and clearly it can be attributed to the actors that committed it, an issue to which we will return later. With these factors in mind, this paper proposes a simple framework that organizes deterrent and defensive options into three broad strategic approaches: *protect and preempt*, *threaten and respond*, and *reassure and reward*. This scheme will look familiar to those acquainted with traditional deterrence theory because it mirrors (with some differences) the coercive trinity of denial, punishment, and positive inducements.[11]

Protect and Preempt

The first strategic approach focuses on defensive measures: making preventive attack appear unlikely to succeed, and therefore not worth the cost of attempting,[12] and if deterrence by denial fails, mitigating the effects

[9] This is intended not to be a definitional straitjacket but something closer to pajamas. It is explicitly looser than standard definitions of preventive war per se, which fundamentally amount to a state going to war because fighting now appears more promising than the expected alternative of fighting later; see Karl P. Mueller, Jasen J. Castillo, Forrest E. Morgan, Negeen Pegahi, and Brian Rosen, *Striking First: Preemptive and Preventive Attack in U.S. National Security Policy*, RAND Corporation, MG-403-AF, 2006.

[10] The specific potential target arrays and attack surfaces presented by the U.S. and Chinese AI ecosystems would of course differ in many respects.

[11] Thomas W. Milburn, "What Constitutes Effective Deterrence?" *Journal of Conflict Resolution*, Vol. 3, No. 2, June 1959; Michael J. Mazarr, *Understanding Deterrence*, RAND Corporation, PE-295-RC, April 2018.

[12] It is important to note that while deterrence by denial emphasizes reducing the probability of an attack succeeding rather than making it costly, it still depends on attacking not being cheap—if it is, there is little incentive not to attack even if the chances of success are miniscule. In conventional warfare, almost everything is inherently expensive, but some of the ways in which AI targets might be attacked involve very small investments of resources on the attacker's part.

of the attack.[13] Reducing prospects for success could entail any of a wide variety of measures (and quite likely a combination of them), including

- active defenses to interdict attacks so they do not reach their targets
- concealment and deception measures to prevent the attacker from aiming at the intended target effectively
- hardening targets to make them more resistant to damage (noting that making a target "harder" is very different when protecting a data center from a missile, software from a hacker, or a worker from assassination or corruption)
- increasing the resilience of a target set through redundancy (including among allies) and capacity for reconstitution and not creating "single point of failure" vulnerabilities to begin with.

A prominent difference between this approach and those that follow is that the measures likely to be employed under this rubric are predominantly unconditional. Instead of threats or promises about what the United States will do in the future depending on the adversary's behavior, the measures involve actions undertaken in advance whose effects would be "baked in" to an adversary's decision about attacking. This leads to a principal drawback with this approach: One mostly pays the costs up front, and while some defenses are relatively inexpensive to implement, many are not, either in terms of direct costs or efficiency losses from heightened physical and cyber security measures, duplication of investments, and the like (and the more valuable one expects developing AGI first to be, the more-serious impediments to AI progress will appear). A second problem is closely related: An AI ecosystem can be attacked at many points and in many ways, so there is likely to be a great deal to protect, and adopting a defensive scheme that leaves some vulnerabilities unshielded may have little value since an adversary can choose to attack the weak points that remain.

This approach need not be entirely defensive in orientation, however. The strategic logic can also extend to taking action to reduce the adversary's offensive capabilities before they can be used or potentially even before they can be developed—essentially mounting preventive attacks of one's own to diminish the threat.[14] In the case of the United States and China, this would of course involve risks of escalation, particularly but not only when conducting attacks at the upper end of the extremity scale, including the possibility of directly or indirectly triggering a major war.

Threaten and Respond

The second strategic approach is to make launching preventive attacks appear too costly to be worthwhile. This option might appeal to leaders if physically protecting against the full range of threats appears either infeasible or unaffordable. In some contexts, such as deterring conventional invasions, threatening to impose prohibitive costs on an attacker on the battlefield may be possible.[15] However, preventive attacks targeting AI would most likely employ means that do not involve hazarding large numbers of enemy troops, such as cyber

[13] See Glenn H. Snyder, *Deterrence and Defense: Toward a Theory of National Security*, Princeton University Press, 1961. Because of the synergy between defense and deterrence by denial, *defensive deterrence* would arguably be a better and less cumbersome name for the latter.

[14] When a defensively motivated attack is launched in order to strike before the adversary does, rather than to fight sooner rather than later, it is most accurately described as *preemptive* rather than *preventive*, but the two categories have much in common. A notable difference between them is that preemptive war may be legal; a merely preventive one usually is not. See Mueller et al. (2006).

[15] John J. Mearsheimer, *Conventional Deterrence*, Cornell University Press, 1983.

warfare, actions by intelligence operatives or special operations forces (or expendable proxies), information operations, or missile strikes. Therefore, threatening to punish an attacker would likely depend on retaliation against the attacking state itself and moreover on threatening greater harm than merely damaging the adversary's AI sector in a tit-for-tat response despite the appealing symmetry of response-in-kind punishment.[16]

Many protect-and-preempt measures, like hardening AI infrastructure and building defenses to protect it, would entail considerable expense and significant opportunity costs (potentially including impeding one's own AI progress). However, the United States already invests enormous resources in maintaining a military that is designed to deter PRC aggression by being able to wage war against China successfully, albeit in response to more traditional geopolitical threats, while states less powerful than their rivals face an uphill climb when seeking to defend themselves. Therefore, relying on punitive threats for deterrence may appear—and indeed may be—a more cost-effective strategic approach, as has often been the case in the past for states facing other types of threats that are difficult to defend against, notably nuclear weapons. The principal challenge that this approach must overcome is one of credibility.

Retaliatory threats typically involve credibility issues if executing the threatened punishment is costly. For nuclear threats, the central solution is that the weapons' destructiveness tends to offset doubts about whether they would be used because even a small chance of catastrophic loss is likely to be taken seriously. The United States' nuclear arsenal, as well as its conventional military, may do much to discourage large-scale preventive attacks against U.S. infrastructure; crippling the extensive American AI ecosystem with kinetic attacks would require more than a few low-casualty pinpricks. Much the same can be said of China, notwithstanding the smaller size of its nuclear force. However, it seems certain that there is some threshold below which more subtle, limited, or ambiguous attacks could be mounted with little or no potential to trigger a nuclear response, although where this lies is likely to be less than certain in advance—and not only to the prospective attacker.

More limited punishments need to be more credible, leading to measures such as making rhetorical commitments that are politically costly to abandon or placing assets that might be attacked near civilian populations to make escalation appear more likely if they are struck.[17] Drawing explicit "redlines" specifying what actions by an adversary will trigger retaliation and what form it will take can also enhance credibility. However, doing so may also enable an opponent to approach the redline in safety without quite touching it when less precise threats might have discouraged going anywhere near it. Moreover, specific and rigid retaliatory commitments tend to be unappealing to leaders who value maintaining room for maneuver, and the past decade has provided abundant examples of U.S. and Russian leaders choosing not to follow through on such threats in the breach. However, it is worth noting that not all retaliatory threats are tempting to renege on even without taking steps to bolster their credibility—in some cases, carrying them out will be a simple act of self-interest in the wake of an adversary attack.[18]

The possibility of limited-scale preventive attacks raises a final, critical issue for this strategic approach: attack attribution. While it is safe to presume that no state could launch an aerial bombardment of U.S. AI

[16] A state that is losing the race to advanced AI might be quite willing to accept an exchange in which it trades its rook for an opponent's queen. See Iskander Rehman, Karl P. Mueller, and Michael J. Mazarr, "Seeking Stability in the Competition for AI Advantage," commentary, RAND Corporation, March 13, 2025. Retaliatory responses that are not in-kind may face greater credibility challenges than more symmetric deterrent threats of punishment, however, if they involve escalatory steps that the adversary doubts one's willingness to take.

[17] For the seminal discussion of credibility management in coercion, see Schelling (1966). The ultimate measure to enhance retaliatory threat credibility is perhaps particularly salient in a discussion focusing on AI: automating the response without a human in the loop, as in Herman Kahn's infamous doomsday machine.

[18] Kenneth A. Oye, *Economic Discrimination and Political Exchange: World Political Economy in the 1930s and 1980s*, Princeton University Press, 1993, Ch. 3.

facilities anonymously (both because of U.S. sensor capabilities and because very few states have the ability to mount such an operation), the picture could be very different for lower-key attacks. An adversary that believes its preventive attack has a good prospect of remaining covert or clandestine may be very difficult to deter with retaliatory threats, a problem that is familiar in the realm of cyber deterrence.[19] On the other hand, the attacks that are small or subtle enough to be readily concealed or denied may be more limited in their effects, which will affect the expected value of conducting them.

Reassure and Reward

There is a third strategic approach option for discouraging preventive attacks: using positive incentives to make not attacking appear more attractive.[20] Although reassurance and rewards to make aggression seem worse than the status quo are typically treated as separate from deterrence, they cannot be separated.[21] Decisions to attack are fundamentally a matter of choosing between those alternatives, as illustrated by cases like Japan in 1941 in which strategic desperation has motivated states to launch disastrous wars. It is plausible to imagine that PRC leaders facing the prospect of a decisive loss in the race to AGI might feel similarly.

The United States should certainly want the rest of the world to view the prospect of U.S. AI leadership and even AGI monopoly with warm enthusiasm. Indeed, if such a view prevailed in Beijing, it would solve the AI preventive action problem for the United States, at least with respect to China. At present this would be an ambitious aspiration, although AI prognosticators should be wary of assuming that the current atmosphere of intense death-stare rivalry between the two powers is destined to continue indefinitely). Thus, U.S. strategists may find that "reassure and reward" is a more promising approach for dealing with other players in the AI arena than for China, and Beijing may feel the same way about the United States, at least in the near term.

Nevertheless, competition does not preclude cooperation, so as we look toward options for establishing a reasonably stable regime on the path to AGI or beyond it, considering how a reassurance and reward approach might work between the United States and China is worthwhile. The spectrum of possibilities far exceeds what can be considered here, but it might include policy options ranging from mechanisms for collaborative AI development and technology exchange between U.S. and Chinese firms,[22] to transparency regimes that allay suspicions about U.S. imperial ambitions, to deliberately making U.S. AI infrastructure vulnerable to preventive attack as a reassurance measure.[23]

In doing so, at least three rules of thumb are worth keeping in mind, however. First, reputation matters in international affairs, and it does not change quickly or easily. Appearing nonthreatening is difficult if it conflicts with existing impressions, since foreign audiences will almost always be able to focus on actions or words that fit established suspicions about U.S. motives or intent. When actions and words do not align, one should expect actions to dominate. Racing hard to win the AI race may make reassurance more difficult, but this does not mean that restraint will necessarily shift impressions very far in the other direction. Second, being vulnerable to punishment for one's future actions can be reassuring to rivals, but this depends

[19] See, for example, Martin C. Libicki, *Cyberdeterrence and Cyberwar*, RAND Corporation, MG-877-AF, 2009.

[20] David A. Baldwin, "The Power of Positive Sanctions," *World Politics*, Vol. 24, No. 1, October 1971.

[21] Milburn (1959) aptly labeled rewards and reassurance as *positive deterrence*, but unfortunately the term never caught on.

[22] Even a multinational collective that excluded China might potentially reassure Beijing if it appeared that U.S. allies would exert a moderating influence on perceived U.S. malign intentions.

[23] Dan Hendrycks, Eric Schmidt, and Alexandr Wang, "Superintelligence Strategy: Expert Version," *SuperIntelligence—Robotics—Safety and Alignment*, Vol. 2, No. 1, March 15, 2025a.

on rivals believing that they will be able to inflict sanctions severe enough to deter one from misbehaving.[24] Mutual vulnerability in the nuclear arena—driven far more by the laws of physics making missile defense difficult than by any U.S. enthusiasm for being held hostage—may represent a unique dynamic because a few nuclear warheads can do so much damage. Yet AI may help make societies vulnerable enough to echo nuclear dynamics on some level. Finally, positive sanctions are not immune to credibility problems.[25] Peace agreements and territorial settlements can be abandoned, mutually beneficial economic exchange can be severed, and benign intentions can turn malignant. Thus, positive sanctions tend to work best when benefits are ongoing but conditional on good behavior and agreements are engineered to minimize incentives to renege.

When dealing with a powerful rival where there is limited room for error, it is important to assemble a coherent strategy with conscious priorities and components that do not work at cross-purposes.

Hybrid Approaches

These three strategic approaches have large areas of practical overlap. Indeed, it would be difficult for either the United States or China to pursue any in isolation even if it wished to do so. It is both natural to seek to combine elements of all of them and sensible to do so, since each seeks to influence a different component of the deterrence calculus in which the expected costs and benefits of both conflict and the status quo are simultaneously in play.

When dealing with a powerful rival where there is limited room for error, it is important to assemble a coherent strategy with conscious priorities and components that do not work at cross-purposes (or at least not more than is unavoidable). Across these approaches there are significant points of incompatibility where choices must be made. One cannot maximize both protection to deter and vulnerability to reassure. Sharing technology and being transparent in its development may reduce rivals' suspicions and hostility but will also tend to increase their capabilities to attack. Executing punitive threats involving military force requires the ability to attack effectively, potentially generating a security dilemma. And of course, attributable preventive actions taken against an enemy's military capabilities, including its military-related AI, may do much to encourage the impression that you are a security threat that needs to be reduced, or indeed that you are preparing to go to war against them.

At the same time, more than a few measures that play important parts in one of these strategic approaches do not appear to do much to undermine the others, making them conspicuously attractive for U.S. strategy. For example, compared with the tools of conventional warfighting, most of the measures that protect one's AI ecosystem in the event of attack do not provide much offensive military power to make a rival insecure. Many resilience measures pose little or no threat to others, although they may entail significant efficiency losses for the state or companies adopting them. Capabilities for attack attribution in the cyber world and other shadowy settings appear to be unequivocally desirable. Many reassurance measures that are not based

[24] Deliberately making the U.S. AI ecosystem sufficiently vulnerable for such a reassurance mechanism to work would not be a trivial problem in physical terms given its scale and the redundancy of many of its elements—or politically given that it principally resides in multiple private-sector firms.

[25] Reid B. C. Pauly, "Damned If They Do, Damned If They Don't: The Assurance Dilemma in International Coercion," *International Security*, Vol. 49, No. 1, Summer 2024.

on creating technological weakness or physical vulnerability do not appear to significantly undermine the deterrent potential of the other approaches.

From Deterrence to Stability

Can we expect these strategic approaches, either largely alone or in more elaborate combination, to work? If deterrence is robust, then international stability may not suffer even if racing toward AGI fosters national insecurities. We know from historical experience that the world is full of successful deterrence. Animus among states and leaders has always been plentiful, yet interstate wars are exceptional events, particularly during the past 80 years. When major wars occur, they are typically attributable in no small part to clever leaders with access to reasonably good information making poor decisions. It does not seem likely that AI will solve this problem.[26] At the same time, however, it is not clear why deterrence should be less effective at preventing wars over issues arising from AI unless leaders believe that AI makes the world much more dangerous. It does appear reasonable to expect AI to accelerate the rate of change in the world, which may in turn (along with AI's novel characteristics) make events less predictable, but since leaders tend to be averse to knowingly taking large risks when they go to war, that may not be a severe problem.

In general, major wars among great powers are likely to continue being the easiest to deter because their costs are clearly high. More-limited actions may appear safer, less expensive, or easier to conceal, complicating the problem of preventing them. Since a small war can turn into a big one, and can grow out of something still smaller, deterring small or subtle preventive attacks is a matter of considerable importance for the strategist. So is resisting temptation to undertake clandestine preventive gambits against a serious adversary without being confident that doing so is worthwhile and that one is prepared to deal with the consequences if the action is exposed or escalation ensues.

As a starting point, there are several policy initiatives that appear worth attempting:

- **Minimize apparent vulnerabilities that might invite preventive attack by a major adversary that is uneasy about your progress toward AGI.** This will be easiest for the United States insofar as the size and diversity of its AI ecosystem already make it difficult to attack effectively. However, a powerful private sector over which the government has only limited influence is well placed to resist making the efficiency sacrifices that often accompany improving resilience if it is not convinced of the need for such measures. Serious red teaming of adversary attack options is called for because the most consequential vulnerabilities might not be the most obvious ones. Being able to identify and attribute clandestine attacks against one's AI efforts is a key element of building such robustness.
- **Avoid overselling the game-changing military and coercive potential of AGI if you are leading the race and your rival is insecure about it.** The stability implications of AGI, like deterrence dynamics more generally, are fundamentally driven by expectations and beliefs. Unfortunately, even if it sought to moderate expectations, the U.S. government (for whom this admonition is most relevant) has a relatively small voice in shaping perceptions about AGI compared with companies and individuals who have potent incentives to do the opposite in order to encourage investment in the sector.
- **Engage in Sino-U.S. dialogue seeking areas of common interest regarding AI risk reduction and mutually beneficial AI use** to help in at least a small way to make the status quo more palatable to both rivals. Remarkably, much advocacy regarding AI predicts breathtaking rates of technological and eco-

[26] Reliably predicting state behavior regarding decisions to go to war is likely to be among the most difficult problems for AGI to crack because of the multiplicity of factors in play and the limited amount of relevant historical data to mine.

nomic change while assuming that the current state of intense rivalry and animosity between Beijing and Washington will remain unaltered over the long term. This is certainly possible (although we know there will be leadership changes in both countries in the foreseeable future, likely before there is AGI), but it should not form the basis of national strategy for the AI competition.

CHAPTER 6

Competition and Collusion: How the AI Arms Race Can Motivate Governance

Jane Vaynman and Tristan A. Volpe

The transformative potential of AI lies in its dual-use nature: The same technology that promises economic revolution could also enable military dominance. Algorithms that optimize global supply chains can coordinate swarms of autonomous drones on the battlefield. This duality explains why AI stands to become the centerpiece of great-power competition, with the United States, China, and other nations already racing to harness its benefits across both civilian and military realms.

The prospects for controlling AI competition are dim. The technology exists in a "dead zone" for arms control—a space where civilian and military applications are virtually indistinguishable and deeply integrated into both sectors. This creates incentives for deception, as states can easily disguise military AI developments behind civilian activities. Verifying compliance with any limitations would require intrusive monitoring, yet such inspections risk exposing proprietary algorithms, training data, and deployment patterns—potentially revealing military vulnerabilities and industrial secrets that adversaries could exploit. As civilian and military AI applications continue to converge, most governance approaches face formidable obstacles.[1]

Effective governance requires shifting our focus from AI as a stand-alone "superweapon" to its fundamental role as an enabling technology that transforms existing systems. Just as the internal combustion engine and digital computing birthed mechanized and cyber warfare, AI could affect almost every modern weapon system, from conventional platforms to space systems. This integration raises a crucial question: Does AI make it easier or harder to distinguish between civilian and military technologies? In some cases, AI erases the physical and functional markers that separate commercial and defense technologies—autonomous attack drones have already become scarcely distinguishable from their peaceful counterparts. In other domains, however, deliberate policy choices—such as requiring military AI systems to use distinct hardware or safety protocols—could draw sharper lines between civilian and military capabilities.

The primary AI leaders—both states and major corporations—share a surprising incentive to maintain clear boundaries between military and civilian AI applications, as distinguishability makes military capabilities more observable, reduces deception incentives, and preserves options for mutual restraint. Rather than pursuing elusive global treaties, the most plausible path forward lies in "AI cartels": coalitions of leading states and companies that establish common standards for the adoption of AI into weapon platforms. AI companies also face strong commercial incentives to voluntarily participate in governance regimes. Historical precedent reveals companies joining restrictive groups to capture economic premiums, avoid exclusion from lucrative markets, and access valuable state-provided resources that reduce compliance costs. By embedding technical markers or operational protocols in military AI systems, cartel members could prevent

[1] On how technology creates this dead zone for cooperation, see Vaynman and Volpe (2023).

AGI's inherent dual-use nature stems from its adaptability to perform virtually any cognitive task, enabling both civilian applications (e.g., economic optimization) and military ones (e.g., autonomous warfare strategy) using the same underlying systems.

rivals from blurring the lines between peaceful and military uses, making collusion between leading AI powers the key to balancing innovation with stability.

Defining the AI Landscape

Our analysis rests on several core assumptions about AI and the geopolitical context in which competition over this technology unfolds.

We define *AGI* as highly autonomous AI capable of matching or surpassing human cognitive capabilities across most economically valuable work—distinct from today's narrow AI systems that excel only in specific predefined functions (e.g., chatbots, image recognition).[2] AGI represents a potential qualitative leap rather than incremental improvement, with transformative impacts on military and economic power. Our timeline for AGI assumes it could emerge within five to ten years, placing arrival around the early to mid-2030s. While our analysis does not depend strongly on this timeline assumption, it suggests a narrow but nontrivial period of time available to states to prepare for increasingly advanced AI systems becoming available for both civilian and military applications. The emergence of AGI would not necessarily close off the opportunities we identify, and may indeed enhance pressures to pursue them, but it may significantly raise uncertainty about the costs and benefits of various forms of competition and cooperation.

AGI's inherent dual-use nature stems from its adaptability to perform virtually any cognitive task, enabling both civilian applications (e.g., economic optimization) and military ones (e.g., autonomous warfare strategy) using the same underlying systems. On the military side, we assume that the technology would create advantages in defense planning, weapon optimization, and battlefield decisionmaking. As a result, our approach considers that AGI will emerge as a "general purpose" technology: a foundational innovation—like electricity or the internet—that enables widespread advancements across multiple sectors by adapting to diverse applications rather than serving a single specialized role.[3]

Our analysis applies to both current AI advancements and potential future breakthroughs like AGI. This approach rests on two key premises. First, we expect no substantial time gap between when major powers achieve significant AI milestones. While a decisive first-mover advantage could theoretically create a window for reshaping the international order, we anticipate that actual developments will occur within time frames too narrow to create such opportunities. Second, even temporary technological leads would be accompanied by considerable uncertainty about their extent and durability, reducing incentives for dramatic revisionist actions.

Recent developments in AI competition between the United States and China support these assumptions. Despite export controls and technological restrictions designed to slow diffusion, these measures

[2] See Ben Goertzel, "Artificial General Intelligence: Concept, State of the Art, and Future Prospects," *Journal of Artificial General Intelligence*, Vol. 5, No. 1, December 2014.

[3] Jeffrey Ding and Allan Dafoe, "Engines of Power: Electricity, AI, and General-Purpose, Military Transformations," *European Journal of International Security*, Vol. 8, No. 3, August 2023. On general-purpose technologies, see Jeffrey Ding, "The Rise and Fall of Technological Leadership: General-Purpose Technology Diffusion and Economic Power Transitions," *International Studies Quarterly*, Vol. 68, No. 2, June 2024b; Jeffrey Ding, *Technology and the Rise of Great Powers: How Diffusion Shapes Economic Competition*, Princeton University Press, 2024a; Richard G. Lipsey, Kenneth I. Carlaw, and Clifford T. Bekar, *Economic Transformations: General Purpose Technologies and Long-Term Economic Growth*, Oxford University Press, 2005.

appear insufficient to fundamentally alter the trajectory toward rough technological parity. China's rapid advancement—exemplified by models like DeepSeek competing effectively with Western counterparts on key benchmarks—illustrates this ongoing convergence.[4] We also assume that breakthroughs toward AGI will be relatively transparent to other companies and governments. The development of AI has remained predominantly commercial with high visibility into technological advances. We expect that powerful economic incentives and continued leadership by private entities will ensure that breakthroughs maintain significant publicity rather than becoming subject to comprehensive classification and secrecy.[5] However, this observable front on the commercial side could be exploited to cover up military applications, which could breed suspicion, as the next section considers in detail.

Our focus is on peacetime dynamics, specifically the way that AI will shape the way that nations build up or limit arms. On the geopolitical front, we therefore assume ongoing competition between the United States and China, characterized by military modernization and preparedness for potential flash points, but without presuming inevitable conflict. This approach allows us to focus on AI's inherent properties as a dual-use enabling technology rather than contending with hypothetical shifts in the balance of power from conflict.

The Problem: AI in the Arms Control Dead Zone

Traditional arms control approaches are widely seen as unworkable for AI, yet the specific reasons remain underexplored. What makes AI uniquely resistant to governance? While some cite uncertainty about its applications or its general-purpose nature, these explanations fail to distinguish AI from other dual-use technologies—such as nuclear or chemical weapons—that have been successfully regulated through international agreements.[6]

The critical distinction lies in AI's specific dual-use attributes, which place it in what we term the *dead zone* for arms control: a realm where civilian and military uses are so deeply integrated that verification becomes impossible without exposing sensitive economic or military secrets.[7] Unlike nuclear technology, which occupies a narrow niche with observable infrastructure, AI's pervasive nature creates insurmountable barriers to cooperation, regardless of political conditions.

Our previous research demonstrates that technologies vary significantly in how their dual-use characteristics affect arms control possibilities. Some technologies, such as aircraft and ships, maintain clear boundaries between military and civilian applications, presenting minimal barriers to verification. Others—like space systems or cyber capability models—blur these boundaries and permeate multiple systems, creating severe information challenges.

In this dead zone, states face a nearly insurmountable dilemma: Military and civilian AI applications appear virtually identical, making it difficult to verify compliance with any agreement without extensive

[4] Tye Graham and Peter W. Singer, "To China, DeepSeek Is More Than an App—It's a Strategic Turning Point," *Defense One*, February 21, 2025.

[5] See, for example, Anthropic's public announcement in November 2024 that it had deployed the Claude frontier model in a Top Secret environment to help the National Nuclear Security Administration (NNSA) evaluate risks. See Sam Sabin, "Exclusive: Anthropic, Feds Test Whether AI Will Share Sensitive Nuclear Info," Axios, November 14, 2024.

[6] For useful starting points, see Baker (2023) and Megan Lamberth and Paul Scharre, "Arms Control for Artificial Intelligence," *Texas National Security Review*, Vol. 6, No. 2, Spring 2023. One notable exception is Hickey, who unpacks how dual-use distinguishability varies across development and deployment stages of AI foundation models (Alan Hickey, "The GPT Dilemma: Foundation Models and the Shadow of Dual-Use," arXiv, arXiv:2407.20442, July 29, 2024).

[7] Vaynman and Volpe, 2023. On the tension between transparency and security as a core information problem for arms control, see Coe and Vaynman (2020).

access to internal systems and code. Yet granting such access risks exposing sensitive information about broader military capabilities and economic assets—arms control inspections thereby present unacceptable security concerns. This creates powerful incentives to double down on competition even when mutual benefits from restraint might exist.[8]

Proposals for an "IAEA for AI"—modeled after the IAEA's role in nuclear governance—are fundamentally flawed when applied to foundational AI models. Unlike nuclear technology, which exists in a relatively isolated niche with dedicated facilities that can be inspected without revealing broader military capabilities, AI is deeply integrated across civilian and military domains. This critical difference explains why states accepted intrusive nuclear inspections: The security risks were largely contained to the four corners of atomic energy enterprises. IAEA access to nuclear power plants, for example, revealed little about conventional military power.

For AI foundation models, however, effective verification mechanisms would likely necessitate the disclosure of algorithms, underlying training datasets, and implementation strategies—a level of transparency that could potentially undermine both national security systems and competitive economic advantages.[9] In this environment, the security costs of transparency inevitably outweigh the benefits of cooperative restraint. The dual-use features of AI technology therefore push states away from arms control toward competitive approaches that keep military assets hidden.

The Pivot: AI's Role in Shaping Distinguishability

With efforts to control AI itself doomed to fail, governance efforts should shift focus from treating AI as a stand-alone "superweapon" to recognizing its fundamental role as an enabling technology that transforms existing systems. AI is not a discrete weapon system with narrowly defined military functions; rather, it is a general-purpose innovation poised to enhance every capability it touches—much as electricity revolutionized both civilian and military domains in the late 19th century. This perspective raises a key question: Will AI make it easier or harder to distinguish between civilian and military technologies?

Our comprehensive analysis of modern weapon technologies reveals a critical pattern: Enabling capabilities can either sharpen or blur this line between military and civilian applications.[10] Some, like the internal combustion engine or stealth technology, made military platforms more distinct from their civilian counterparts. Others, such as digital computing and advanced manufacturing, blurred the boundaries between civilian and military capabilities. Understanding how AI will shape this distinguishability dynamic is central to today's great-power competition. States are not merely racing to develop advanced AI in isolation but are competing over how best to use AI to enhance their existing weapon technologies. The ultimate concern is whether this enhancement will push more capabilities into the arms control dead zone, create new opportunities for restraint, or drive new forms of arms racing.

When weapon technologies are enhanced by AI, the distinguishability between civilian and military uses can shift in two ways: technical and political. The next section specifies how technical effects will likely push

[8] See Jane Vaynman and Tristan A. Volpe, "Duplicity and Disclosure: How Technology Shapes Arming Strategies," workshop, Stanford University, June 6, 2024.

[9] On efforts to manage this disclosure problem with technical solutions, see Brundage et al. (2020); and Matthew Mittelsteadt, *AI Verification: Mechanisms to Ensure AI Arms Control Compliance*, Center for Security and Emerging Technology, February 2021. See also Michael C. Horowitz and Lauren A. Kahn, "Nuclear Non-Proliferation Is the Wrong Framework for AI Governance," AI Frontiers, June 27, 2025.

[10] We benchmark dual-use distinguishability across all major weapon technologies available to states over the last 150 years; see Vaynman and Volpe (2023).

AI-enabled weapon systems toward lower distinguishability. This sets the stage to consider political interventions to sharpen the line between military and civilian capabilities.

Technical Erosion

History illustrates how technological advancements can erode the line between civilian and military systems through changes in physical features. The evolution of space technology is instructive: Over time, civilian and military satellites converged in capabilities, making it increasingly difficult to differentiate peaceful uses from espionage or weapon platforms.[11] AI appears poised to accelerate this trend across multiple domains, fundamentally altering two critical dimensions of distinguishability.

First, AI dramatically accelerates the conversion of civilian capabilities into military tools. Commercial AI systems—such as logistics algorithms optimizing supply chains, autonomous drones delivering packages, or facial recognition software managing security—can be rapidly repurposed for warfare. For example, a civilian drone network designed for agricultural monitoring could, with minimal adjustments, deploy swarms for battlefield reconnaissance or strikes. This agility stems from AI's reliance on software and data rather than specialized hardware, enabling militaries to co-opt civilian systems at unprecedented speed. During the 2020 Nagorno-Karabakh conflict, commercially available drones were swiftly weaponized, underscoring this vulnerability.[12] With AI, such conversions could occur in hours rather than months, as machine learning models retrain for new tasks.[13]

Second, AI erodes the physical and functional features that traditionally distinguished military systems. Autonomous weapons, for instance, may no longer require crew compartments, life-support systems, or other human-centric design elements that once made warships or aircraft identifiable. A civilian cargo vessel equipped with AI navigation could be retrofitted with missile launchers and rerouted as a stealthy, crewless warship, indistinguishable from its commercial counterparts. Similarly, AI-driven cyber tools—whether designed for network defense or corporate data analysis—share underlying architectures that blur the line between civilian and offensive capabilities. Even in domains like biotechnology, AI models developed for drug discovery could be repurposed to engineer pathogens, with no outward signs of militarization at the laboratory.

While exceptions may exist—such as AI systems developed specifically for niche military applications—the broader trend is troubling. Unlike many past technologies that required visible, time-intensive efforts to weaponize, AI enables greater secrecy and ambiguity. For military technologies where military and civilian applications are already difficult to distinguish, such as biotechnology and cyber, the integration of AI further exacerbates rather than mitigates the problem. The more critical question is how AI will affect technologies that currently are quite distinguishable, including many conventional weapon platforms. The technical pattern suggests that, without political intervention, AI-enabled civilian and military capabilities will be significantly harder to distinguish than their non-AI predecessors.

[11] For detailed historical examinations of this convergence, see Aaron Bateman, *Weapons in Space: Technology, Politics, and the Rise and Fall of the Strategic Defense Initiative*, MIT Press, 2024; and Deganit Paikowsky, "Dual Use of Space Technology: A Challenge or an Opportunity? Space Commercialization in the US After the Cold War," in Brian C. Odom, ed., *The Rise of the Commercial Space Industry: Early Space Age to the Present*, Springer, 2024.

[12] Shaan Shaikh and Wes Rumbaugh, "The Air and Missile War in Nagorno-Karabakh: Lessons for the Future of Strike and Defense," Center for Strategic and International Studies, December 8, 2020.

[13] On conversion speed dynamics with AI foundation models, see Hickey (2024).

Political Enhancements

The erosion of boundaries between civilian and military AI applications is not inevitable—it can be countered through deliberate policy choices. This political pathway to preserving distinguishability may prove more consequential than technical factors, particularly if states act early in AI's development to establish norms and standards that clearly differentiate civilian from military applications. These choices will shape whether AI-enabled weapon technologies remain observable and verifiable or slip into the arms control dead zone.

Several historical cases reveal how policy interventions, often made for entirely different reasons, have profoundly reshaped the distinguishability of technologies. Consider the deployment decisions for long-range rockets, specifically intercontinental ballistic missiles (ICBMs) and space launch vehicles (SLVs), made by the United States and the Soviet Union during the Cold War. Both capabilities started with the same deployment patterns in the 1960s.

But policymakers soon came to prioritize military survivability for ICBMs, deliberately placing hardened missile silos underground at remote ranges by the 1970s. Simultaneously, civilian space authorities made the calculated decision to optimize SLVs for payload capacity, maintaining above-ground, liquid-fueled systems. These deliberate policy choices, rather than inevitable technological evolution, created the readily observable distinctions between peaceful and military applications that we recognize today.

Today, the primary AI leaders—both nations and major corporations—face decisions that will shape the future distinguishability of military systems. Security requirements could mandate different architectural approaches for integrating AI into weapon systems, such as using specialized training data, distinct algorithms, or unique operational protocols. These measures would create development pathways separate from those of civilian applications.

Similarly, requiring AI used in weapon systems to incorporate verifiable safety measures or technical "watermarks" absent in civilian models could establish observable differences between military and civilian systems.[14] Even basic deployment choices, such as operating military AI on isolated networks with distinct hardware, could generate signatures detectable by monitoring regimes. Conversely, if the AI community broadly accepts systems with low explainability—such as opaque deep learning models that cannot clarify why they identify objects in satellite imagery as military targets versus civilian infrastructure—this could lead to convergence in development standards and significantly reduce barriers between military and civilian applications.

The choices made around AI development and deployment—often driven by immediate security or commercial concerns—will play a key role in determining whether AI-enabled military systems remain observable and potentially subject to verification. Cruise missiles and other long-range standoff munitions, for example, have long been highly distinguishable from their civilian cousins. Some of these weapons are already adopting more-advanced AI targeting systems that can make independent targeting decisions, coordinate swarm formations, and plan routes around defensive systems—potentially blurring the boundaries with small commercial drones that leverage similar breakthroughs in autonomy.[15] However, a more deliberate approach is possible. States and industry leaders can actively seek to influence the distinguishability of weapon technologies, recognizing its strategic value.

---

[14] Shoker et al., 2023; Robert F. Trager, Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, Ranjit Lall, Owen Larter, Seán Ó hÉigeartaigh, et al., *International Governance of Civilian AI: A Jurisdictional Certification Approach*, Oxford Martin Centre for the Governance of AI, 2023.

[15] See, for example, the openly acknowledged AI capabilities onboard the AGM-158C Long-Range Anti-Ship Missiles (LRASMs) that Lockheed Martin currently builds (John Keller, "Air Force Orders Artificial Intelligence [AI]-Enabled Anti-Ship Missiles with Imaging Infrared Sensors," Military and Aerospace Electronics, July 14, 2025).

The Opportunity: Collusion Strategies for AI Governance

The fierce competition over advanced AI may drive leading powers toward selective collusion rather than unrestrained arms racing. While conventional wisdom envisions an inevitable race to develop AI without constraints, the small group of AI leaders—both states and major corporations—share an incentive to maintain clear boundaries between military and civilian applications of AI-enabled technologies.

The main technology leaders benefit from a competitive environment where weapon systems with fresh AI enhancements remain sharply distinct from civilian counterparts and where deception is both difficult and unnecessary for military buildups. When military applications are distinguishable, states can use technology investments to create deterrence leverage in peacetime while maintaining advantages in the event of conflict. Observable boundaries between civilian and military applications also preserve the option to pursue selective arms control agreements when particular capabilities become destabilizing or prohibitively expensive.[16]

By establishing governance frameworks that encourage distinguishability in military technology, AI leaders can create path dependencies that institutionalize transparent development while making deception more costly. The nuclear nonproliferation regime provides an instructive parallel: States that subscribed to Nuclear Non-Proliferation Treaty (NPT) safeguards later faced higher costs when attempting to hide military nuclear programs. Iran encountered significantly greater challenges concealing its nuclear activities precisely because transparency commitments were already established when it began pursuing weapon capabilities. But what is the best governance structure to encourage the adoption of AI in ways that clarify the line between military and civilian applications?[17]

Hegemonic and Duopolistic Efforts

A hegemonic approach—where a single dominant actor sets and enforces the rules—is unlikely to succeed, since AI development is highly diffuse and lacks the choke points found in technologies that depend on rare materials or specialized manufacturing facilities. The United States' recent attempts to restrict China's access to advanced semiconductors illustrate this challenge: Despite these measures, Chinese firms like DeepSeek have continued to develop competitive AI models, suggesting that even comprehensive technological containment strategies have significant limitations in this domain.[18]

Bilateral collusion between leading powers, similar to U.S.-Soviet cooperation on the NPT, also appears insufficient.[19] The AI landscape is already more multipolar, with numerous actors—the United States, China, the European Union, and others—possessing significant capabilities. Unlike nuclear technology, private actors may be as central as states in shaping deployment patterns. Industry leaders like OpenAI, Anthropic, and Google have already demonstrated their interest in governance through their support for safety standards. Some governance structures, while ostensibly promoting responsible behavior, could effectively entrench industry leaders' advantages while creating higher barriers to entry.

---

[16] See Jess Whittlestone and Jack Clark, "Why and How Governments Should Monitor AI Development," arXiv, arXiv:2108.12427, August 31, 2021.

[17] For a complementary argument that effective global governance of AI will require a decentralized web of overlapping institutions and initiatives rather than a single global body, see Emma Klein and Stewart Patrick, *Envisioning a Global Regime Complex to Govern Artificial Intelligence*, Carnegie Endowment for International Peace, March 2024.

[18] Nikita Lalwani, "How America Can Stay Ahead of China in the AI Race: The Case for Export Control Diplomacy," *Foreign Affairs*, April 15, 2025; Lennart Heim, *Understanding the Artificial Intelligence Diffusion Framework: Can Export Controls Create a U.S.-Led Global Artificial Intelligence Ecosystem?* RAND Corporation, PE-A3776-1, January 2025.

[19] On this strategy, see Coe and Vaynman (2015).

Cartel Collusion

A more viable approach may be a broader cartel-like arrangement among the small group of leading AI powers. Cartels function most effectively when membership is limited to producers with similar interests and capabilities, allowing them to coordinate policies and enforce compliance.[20] In the AI context, a cartel could establish two-tier governance systems where privileged members operate under one set of rules while imposing more-restrictive conditions on nonmembers. This mechanism would allow technological leaders to formalize their collaboration while providing collective enforcement against defectors. A cartel would make military AI applications more distinguishable from civilian uses through coordinated technical standards and institutional enforcement mechanisms that maintain transparency where natural technical differences might not exist.

The specific rules and requirements a cartel could adopt to promote distinguishability would be the subject of a future, more extensive study. However, a few preliminary examples illustrate the concept. The cartel could mandate verifiable watermarks embedded directly into military AI systems, functioning like cryptographic identifiers that clearly mark systems as military grade. These digital signatures would be required for all defense applications while remaining optional for civilian systems, creating immediate distinguishability during inspections. Additionally, the cartel could establish distinctive architectural requirements for military AI, such as specialized hardware configurations, dedicated AI chips with military-specific features, or physically air-gapped development environments that isolate military AI development from civilian networks.[21]

Beyond technical identification, military AI systems could be required to maintain human-interpretable decision paths and explainability features that civilian applications would not need. While civilian AI often operates as opaque "black boxes," military systems would include transparency requirements, verification zones for critical functions, and checkpoint systems that make their reasoning processes auditable and distinguishable from civilian counterparts. This transparency requirement would create observable operational differences between military and civilian AI applications.[22]

A cartel would make military AI applications more distinguishable from civilian uses through coordinated technical standards and institutional enforcement mechanisms that maintain transparency where natural technical differences might not exist.

The cartel would extend these requirements through alliance networks, making compliance with distinguishability standards a condition of receiving military AI technology or security guarantees. Similarly to how the Nuclear Suppliers Group (NSG) requires safeguards for civilian nuclear technology, the AI cartel would create cascading effects where even non-cartel members would need to adopt distinguishability measures to access

[20] States that possess militarily significant technologies often create exclusive clubs to mitigate collective challenges posed by the spread of these weapon systems—such as arms races or instability—and to protect their own strategic advantages by controlling who can access, use, or transfer these sensitive capabilities. See Eliza Gheorghe, "Proliferation and the Logic of the Nuclear Market," *International Security,* Vol. 43, No. 4, Spring 2019; and Gadi Heimann, Deganit Paikowsky, and Or Rabinowitz, "Sneaking Through Raising Walls: The Dynamics of Institutionalizing Security Technology Clubs," *Technology in Society*, Vol. 77, June 2024.

[21] See Mittelsteadt (2021) and Baker (2023).

[22] See Brundage et al. (2020) and Shoker et al. (2023).

advanced military AI capabilities.[23] This incentive mechanism would expand the reach of distinguishability standards beyond cartel members themselves.

Finally, the cartel could establish compliance verification mechanisms, including regular audits of military AI systems, monitoring of development environments, and verification that required technical features are present and functioning.[24] By institutionalizing these standards early in AI development, the cartel would create path dependencies that make it costly and technically challenging to remove distinguishability features or develop covert capabilities. This approach effectively transforms what is currently a technical challenge—distinguishing military from civilian AI—into a governance solution maintained through coordinated standards rather than relying on inherent technical differences.[25]

Why Companies Join Governance Regimes: Lessons for AI Cartel Formation

The commercial drivers behind AI development differ from past technologies that were initially developed under government control, such as nuclear weapons. AI development is primarily led by private companies pursuing commercial objectives. Yet those commercial origins do not preclude government involvement. States routinely participate in governance regimes involving private actors, driven by concerns for national security, public safety, and market access for domestic companies. International regimes vary widely in their priorities and government roles. Some focus primarily on security concerns, others on market access, and many have no national security implications at all. Government involvement ranges from setting rules on private actors to providing enforcement for jointly developed standards or assisting with information-gathering and standardization.

Many regimes are hybrid, addressing both market and security priorities, and can evolve over time. The NSG began with economic motivations to ensure fair commercial competition among transnational nuclear enterprises before shifting to include nuclear proliferation concerns.[26] The main member states (e.g., the United States, Russia, France) coordinated rules for civilian nuclear exports among themselves while impos-

[23] See Gheorghe (2019); Rebecca Davis Gibbons, *The Hegemon's Tool Kit: US Leadership and the Politics of the Nuclear Nonproliferation Regime*, Cornell University Press, 2022; and Lisa Langdon Koch, "Frustration and Delay: The Secondary Effects of Supply-Side Proliferation Controls," *Security Studies*, Vol. 28, No. 4, 2019.

[24] See Heim (2025) and Mittelsteadt (2021).

[25] Competitive states will have incentives to cheat on any agreement. The viability of creating governance regimes depends on the conditions under which incentives to cheat are sufficiently low, and their detrimental effects can be sufficiently mitigated through careful agreement design. We can consider at least two critical conditions: (1) the period during which any leader can maintain a secret advantage must be sufficiently short, as longer time frames create stronger cheating incentives, and (2) the magnitude of any advantage gained through cheating must be limited, since high advantages that can be applied quickly make cheating too dangerous for anyone to risk creating an agreement. A cartel-type regime for AI governance, therefore, becomes possible under specific assumptions, including those about the development of AGI which we noted at the outset: (1) being a first mover on AGI must not provide any single state with massive preemptive advantages to eliminate adversaries, (2) reducing distinguishability for previously highly distinguishable military technologies should not grant states significant or long-term security advantages, and (3) adopting governance rules must not reveal private corporate information that might compromise participants' broader strategic positions or create new vulnerabilities. A significant change in these assumptions could render cheating incentives too high for any actor to agree to cooperation in the first place.

[26] For a detailed account, see Isabelle Anstey, "Negotiating Nuclear Control: The Zangger Committee and the Nuclear Suppliers' Group in the 1970s," *International History Review*, Vol. 40, No. 5, 2018. See also Gheorghe (2019).

ing stricter restrictions on nonmembers, like India, which was barred from accessing sensitive technologies until securing a contested exemption in 2008 after demonstrating compliance with nonproliferation norms.[27]

AI governance would likely follow this hybrid pattern, encompassing both national security objectives for maintaining dual-use distinguishability and market access elements independent of military considerations. The NSG example even suggests that an AI governance regime could similarly begin with commercial coordination before expanding to address security concerns, or alternatively, the reverse, focusing on core national security elements first and expanding to encompass more of the commercial coordination needs.

An "AI cartel" focused on maintaining distinguishability for military applications would require participation from the multinational companies currently leading AI development. Although some AI developers express interest in managing global risks, most are primarily motivated by commercial incentives. This raises a critical question: Why would powerful companies voluntarily join regimes that constrain their technology development and exports?

The historical record reveals that major industry players embrace governance regimes for three primary reasons. First, companies seek to capture market premiums. When sufficient buyers pay premiums for certified goods or exclusively purchase certified products, companies actively establish regimes to capture these advantages. The Forest Stewardship Council (FSC) certification demonstrates this dynamic—FSC-certified wood products command higher prices while major retailers mandate certification as a business prerequisite.

Second, beyond market premiums, companies join to avoid competitive disadvantages. When governments create state-run regimes, companies prefer participation over exclusion even if they would rather have no regime at all. The U.S. chemical industry's Chemical Weapons Convention participation exemplifies this necessity: Companies faced potential $600 million annual losses and reputational damage if excluded from regulated markets. This exclusion fear particularly motivates dominant companies to participate defensively against losing market position to compliant competitors.

Third, companies are incentivized by government-provided resources. When governments provide regime members with valuable benefits that reduce participation costs, companies find joining attractive. The Kimberley Process illustrates how states provide enforcement infrastructure through customs agencies and border controls that companies cannot independently replicate. By coordinating information-sharing systems across 86 countries, governments make regime participation more cost-effective than developing independent compliance systems.

AI companies are likely to find themselves facing similar incentives. Governments are major buyers of AI applications and can set the standards that companies must meet to access government and particularly military customers as lucrative markets. Given governments' strong security motivations, they will likely pursue international AI governance rules regardless of industry preferences, making it advantageous for AI companies to participate in shaping these rules rather than face exclusion from the process. Additionally, governments possess unique capabilities, such as advanced detection tools or enforcement methods, that would make compliance with AI governance regimes more cost-effective than developing independent solutions. These dynamics—market premiums, exclusion avoidance, and government-subsidized resources—suggest that AI companies may voluntarily embrace governance structures that impose constraints on their operations, viewing participation as economically rational rather than merely regulatory compliance. The historical precedent of other technology governance regimes indicates that commercial incentives, rather than altruistic concerns about global risks, are likely to drive industry participation in emerging AI governance frameworks.

[27] Mark Hibbs, *The Future of Nuclear Power in China*, Carnegie Endowment for International Peace, 2018; Koch, 2019.

Back to the Future of Technology Governance

The future of AI governance will likely not be shaped by sweeping global treaties or attempts to control the technology in its entirety but by the strategic choices of a small group of leading states and corporations. As AI blurs the boundaries between civilian and military realms, the risk of an ungovernable dead zone grows—one where verification is impossible and deception is rewarded.

Yet history shows that even in the face of profound technical challenges and strong incentives for competition, institutional innovation is possible.[28] The nuclear nonproliferation regime offers an instructive model for AI governance—not as a blueprint for controlling AI itself but as a framework for how powerful states can manage competition while preserving strategic advantages. The NPT succeeded by creating a system where leading powers could efficiently coordinate their interests while imposing higher costs on potential spoilers.[29] This logic was reinforced by the NSG, which used its cartel-like arrangement to further restrict access to critical nuclear technologies for noncompliant states, thereby amplifying the costs of deception.

This historical parallel suggests a counterintuitive prescription: Effective management of AI may require selective collusion rather than either universal restraint or competition. By embracing cartel-like arrangements that prioritize distinguishability, AI leaders can carve out a path for responsible competition that reduces incentives for deception and allows for more-stable competition in both commercial and military spheres.

The window for establishing these governance structures is likely narrowing. As AI becomes more deeply integrated into military and civilian systems, implementing political interventions to address dual-use concerns will become increasingly difficult. Advanced AI and eventually AGI could usher in an environment where all AI-enabled weapon development occurs in the shadows, obscured by commercial competition. If leading powers recognize their shared interest in avoiding this outcome, embedding transparency mechanisms into AI-enabled weapon systems offers a viable and attractive strategy for pragmatic cooperation—even as broader commercial and geopolitical competition continues.

---

[28] On the most viable governance regimes for AI, see also Klein and Patrick (2024).

[29] Coe and Vaynman, 2015.

Abbreviations

AGI	artificial general intelligence
AI	artificial intelligence
C4	command, control, communications, and computers
C4ISRT	command, control, communications, computers, intelligence, surveillance, reconnaissance, and targeting
IAEA	International Atomic Energy Agency
ICBM	intercontinental ballistic missile
NC3	nuclear command, control, and communications
NPT	Nuclear Non-Proliferation Treaty
NSG	Nuclear Suppliers Group
PRC	People's Republic of China
R&D	research and development
SLV	space launch vehicle
WMD	weapons of mass destruction

References

Altmann, Jürgen, "Verification Is Possible: Checking Compliance with an Autonomous Weapon Ban," *Lawfare* blog, April 8, 2024.

Anderljung, Markus, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety," arXiv, arXiv:2307.03718, November 7, 2023.

Anstey, Isabelle, "Negotiating Nuclear Control: The Zangger Committee and the Nuclear Suppliers' Group in the 1970s," *International History Review*, Vol. 40, No. 5, 2018.

Anthropic, "Anthropic's Responsible Scaling Policy," webpage, last updated May 14, 2025a. As of September 3, 2025:
https://www.anthropic.com/rsp-updates

Anthropic, "Activating AI Safety Level 3 Protections," May 22, 2025b.

Aschenbrenner, Leopold, "Situational Awareness: The Decade Ahead," Situational-Awareness.ai, June 2024.

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

Baldwin, David A., "The Power of Positive Sanctions," *World Politics*, Vol. 24, No. 1, October 1971.

Bateman, Aaron, *Weapons in Space: Technology, Politics, and the Rise and Fall of the Strategic Defense Initiative*, MIT Press, 2024.

Blainey, Geoffrey, *The Causes of War*, 3rd ed., Free Press, 1988.

Brent, Roger, T. Greg McKelvey, Jr., and Jason Matheny, "The New Bioweapons: How Synthetic Biology Could Destabilize the World," *Foreign Affairs*, August 20, 2024.

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al., "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims," arXiv, arXiv:2004.07213, April 20, 2020.

Burdette, Zachary, and Hiwot Demelash, "The Risks of Preventive Attack in the Race for Advanced Artificial Intelligence," RAND Corporation, WR-A4005-1, 2025. As of July 3, 2025:
https://www.rand.org/pubs/working_papers/WRA4005-1.html

Burdette, Zachary, Karl Mueller, Jim Mitre, and Lily Hoak, "Six Ways AI Could Cause the Next Big War, and Why It Probably Won't," *Bulletin of the Atomic Scientists*, July 15, 2025.

Bureau of Arms Control, Deterrence, and Stability, "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," U.S. Department of State, last updated November 27, 2024.

Clymer, Joshua, Nick Gabrieli, David Krueger, and Thomas Larsen, "Safety Cases: How to Justify the Safety of Advanced AI Systems," arXiv, arXiv:2403.10462, March 18, 2024.

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

Coe, Andrew J., and Jane Vaynman, "Why Arms Control Is So Rare," *American Political Science Review*, Vol. 114, No. 2, May 2020.

Crootof, Rebecca, Margot E. Kaminski, and W. Nicholson Price II, "Humans in the Loop," *Vanderbilt Law Review*, Vol. 76, No. 2, May 2023.

Dafoe, Allan, Anca Dragan, Four Flynn, Helen King, Tom Lue, Lewis Ho, and Rohin Shah, "Updating the Frontier Safety Framework," webpage, DeepMind, February 4, 2025. As of September 3, 2025:
https://deepmind.google/discover/blog/updating-the-frontier-safety-framework/

Davidson, Tom, Lukas Finnveden, and Rose Hadshar, *AI-Enabled Coups: How a Small Group Could Use AI to Seize Power*, Forethought Foundation, April 2025.

Ding, Jeffrey, *Technology and the Rise of Great Powers: How Diffusion Shapes Economic Competition*, Princeton University Press, 2024a.

Ding, Jeffrey, "The Rise and Fall of Technological Leadership: General-Purpose Technology Diffusion and Economic Power Transitions," *International Studies Quarterly*, Vol. 68, No. 2, June 2024b.

Ding, Jeffrey, and Allan Dafoe, "Engines of Power: Electricity, AI, and General-Purpose, Military Transformations," *European Journal of International Security*, Vol. 8, No. 3, August 2023.

Drexel, Bill, *Promethean Rivalry: The World-Altering Stakes of Sino-American Competition*, Center for a New American Security, April 2025.

Fearon, James D., "Causes and Counterfactuals in Social Science," in Philip E. Tetlock and Aaron Belkin, eds., *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*, Princeton University Press, 1996.

Fink, Anya L., *Russia's Nuclear Weapons: Doctrine, Forces, and Modernization*, Congressional Research Service, R45861, April 21, 2022.

flexHEG, homepage, undated. As of July 7, 2025:
https://flexheg.com

Geist, Edward, *Deterrence Under Uncertainty: Artificial Intelligence and Nuclear Warfare*, Oxford University Press, 2023.

Geist, Edward, and Alvin Moon, "What Even Superintelligent Computers Can't Do: A Preliminary Framework for Identifying Fundamental Limits Constraining Artificial General Intelligence," RAND Corporation, WR-A3990-1, 2025. As of July 29, 2025:
https://www.rand.org/pubs/working_papers/WRA3990-1.html

Gheorghe, Eliza, "Proliferation and the Logic of the Nuclear Market," *International Security*, Vol. 43, No. 4, Spring 2019.

Gibbons, Rebecca Davis, *The Hegemon's Tool Kit: US Leadership and the Politics of the Nuclear Nonproliferation Regime*, Cornell University Press, 2022.

Goertzel, Ben, "Artificial General Intelligence: Concept, State of the Art, and Future Prospects," *Journal of Artificial General Intelligence*, Vol. 5, No. 1, December 2014.

Graham, Tye, and Peter W. Singer, "To China, DeepSeek Is More Than an App—It's a Strategic Turning Point," *Defense One*, February 21, 2025.

Harack, Ben, Robert F. Trager, Anka Reuel, David Manheim, Miles Brundage, Onni Aarne, Aaron Scher, Yanliang Pan, Jenny Xiao, Kristy Loke, et al., "Verification for International AI Governance," AI Governance Initiative, Oxford Martin School, University of Oxford, July 3, 2025.

Heim, Lennart, *Understanding the Artificial Intelligence Diffusion Framework: Can Export Controls Create a U.S.-Led Global Artificial Intelligence Ecosystem?* RAND Corporation, PE-A3776-1, January 2025. As of August 6, 2025:
https://www.rand.org/pubs/perspectives/PEA3776-1.html

Heimann, Gadi, Deganit Paikowsky, and Or Rabinowitz, "Sneaking Through Raising Walls: The Dynamics of Institutionalizing Security Technology Clubs," *Technology in Society*, Vol. 77, June 2024.

Hendrycks, Dan, Eric Schmidt, and Alexandr Wang, "Superintelligence Strategy: Expert Version," *SuperIntelligence—Robotics—Safety and Alignment*, Vol. 2, No. 1, March 15, 2025a.

Hendrycks, Dan, Eric Schmidt, and Alexandr Wang, "Superintelligence Strategy: Expert Version," arXiv, arXiv:2503.05628, April 14, 2025b.

Hibbs, Mark, *The Future of Nuclear Power in China*, Carnegie Endowment for International Peace, 2018.

Hickey, Alan, "The GPT Dilemma: Foundation Models and the Shadow of Dual-Use," arXiv, arXiv:2407.20442, July 29, 2024.

Horowitz, Michael C., "Autonomous Weapon Systems: No Human-in-the-Loop Required, and Other Myths Dispelled," *War on the Rocks*, May 22, 2025.

Horowitz, Michael C., and Lauren A. Kahn, "Nuclear Non-Proliferation Is the Wrong Framework for AI Governance," AI Frontiers, June 27, 2025.

Horowitz, Michael C., and Paul Scharre, *AI and International Stability: Risks and Confidence-Building Measures*, Center for a New American Security, January 2021.

Huang, Raffaele, and Liza Lin, "Chinese AI Companies Dodge U.S. Chip Curbs by Flying Suitcases of Hard Drives Abroad," *Wall Street Journal*, June 12, 2025.

Jervis, Robert, "Cooperation Under the Security Dilemma," *World Politics*, Vol. 30, No. 2, January 1978.

Kahl, Colin H., and Jim Mitre, "The Real AI Race: America Needs More Than Innovation to Compete with China," *Foreign Affairs*, July 9, 2025.

Keller, John, "Air Force Orders Artificial Intelligence (AI)-Enabled Anti-Ship Missiles with Imaging Infrared Sensors," Military and Aerospace Electronics, July 14, 2025.

Kissinger, Henry A., Eric Schmidt, and Daniel Huttenlocher, *The Age of AI: And Our Human Future*, Back Bay Books, 2022.

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

Koblentz, Gregory, "Pathogens as Weapons: The International Security Implications of Biological Warfare," *International Security*, Vol. 28, No. 3, Winter 2003–2004.

Koblentz, Gregory D., *Living Weapons: Biological Warfare and International Security*, Cornell University Press, 2009.

Koch, Lisa Langdon, "Frustration and Delay: The Secondary Effects of Supply-Side Proliferation Controls," *Security Studies*, Vol. 28, No. 4, 2019.

Kulp, Gabriel, Daniel Gonzales, Everett Smith, Lennart Heim, Prateek Puri, Michael J. D. Vermeer, and Zev Winkelman, "Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090," RAND Corporation, WR-A3056-1, 2024. As of July 3, 2025:
https://www.rand.org/pubs/working_papers/WRA3056-1.html

Lalwani, Nikita, "How America Can Stay Ahead of China in the AI Race: The Case for Export Control Diplomacy," *Foreign Affairs*, April 15, 2025.

Lamberth, Megan, and Paul Scharre, "Arms Control for Artificial Intelligence," *Texas National Security Review*, Vol. 6, No. 2, Spring 2023.

Lee, Kai-Fu, *AI Superpowers: China, Silicon Valley, and the New World Order*, Harper Business, 2018.

Lemmer, Felix, "Poseidon: Oceanic Multipurpose System Status-6, Kanyon," Hertie School Centre for International Security, March 2022.

Libicki, Martin C., *Cyberdeterrence and Cyberwar*, RAND Corporation, MG-877-AF, 2009. As of July 3, 2025:
https://www.rand.org/pubs/monographs/MG877.html

Lim, Andrew S., and James D. Fearon, "The Conventional Balance of Terror: America Needs a New Triad to Restore Its Eroding Deterrence," *Foreign Affairs*, April 22, 2025.

Lin-Greenberg, Erik, "Wars Are Not Accidents: Managing Risk in the Face of Escalation," *Foreign Affairs*, October 8, 2024.

Lipsey, Richard G., Kenneth I. Carlaw, and Clifford T. Bekar, *Economic Transformations: General Purpose Technologies and Long Term Economic Growth*, Oxford University Press, 2005.

Mazarr, Michael J., *Understanding Deterrence,* RAND Corporation, PE-295-RC, April 2018. As of June 26, 2025:
https://www.rand.org/pubs/perspectives/PE295.html

Mearsheimer, John J., *Conventional Deterrence*, Cornell University Press, 1983.

Metz, Cade, "A Hacker Stole OpenAI Secrets, Raising Fears That China Could, Too," *New York Times*, July 4, 2024.

Milburn, Thomas W., "What Constitutes Effective Deterrence?" *Journal of Conflict Resolution*, Vol. 3, No. 2, June 1959.

Mitre, Jim, and Joel B. Predd, *Artificial General Intelligence's Five Hard National Security Problems*, RAND Corporation, PE-A3691-4, February 2025. As of July 8, 2025:
https://www.rand.org/pubs/perspectives/PEA3691-4.html

Mittelsteadt, Matthew, *AI Verification: Mechanisms to Ensure AI Arms Control Compliance*, Center for Security and Emerging Technology, February 2021.

Morrow, James D., *Order Within Anarchy: The Laws of War as an International Institution*, Cambridge University Press, 2014.

Mueller, Karl P., *Heeding the Risks of Geopolitical Instability in a Race to Artificial General Intelligence*, RAND Corporation, PE-A3691-12, July 2025. As of July 25, 2025:
https://www.rand.org/pubs/perspectives/PEA3691-12.html

Mueller, Karl P., Jasen J. Castillo, Forrest E. Morgan, Negeen Pegahi, and Brian Rosen, *Striking First: Preemptive and Preventive Attack in U.S. National Security Policy*, RAND Corporation, MG-403-AF, 2006. As of July 3, 2025:
https://www.rand.org/pubs/monographs/MG403.html

Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott, *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*, RAND Corporation, RR-A2849-1, 2024. As of July 3, 2025:
https://www.rand.org/pubs/research_reports/RRA2849-1.html

OpenAI, "Our Updated Preparedness Framework," webpage, April 15, 2025. As of September 3, 2025:
https://openai.com/index/updating-our-preparedness-framework/

Oye, Kenneth A., *Economic Discrimination and Political Exchange: World Political Economy in the 1930s and 1980s*, Princeton University Press, 1993.

Paikowsky, Deganit, "Dual Use of Space Technology: A Challenge or an Opportunity? Space Commercialization in the US After the Cold War," in Brian C. Odom, ed., *The Rise of the Commercial Space Industry: Early Space Age to the Present*, Springer, 2024.

Pauly, Reid B. C., "Damned If They Do, Damned If They Don't: The Assurance Dilemma in International Coercion," *International Security*, Vol. 49, No. 1, Summer 2024.

Petrie, James, Onni Aarne, Nora Ammann, and David Dalrymple, *Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees*, Institute for AI Policy and Strategy and Advanced Research and Invention Agency, August 23, 2024.

Rehman, Iskander, Karl P. Mueller, and Michael J. Mazarr, "Seeking Stability in the Competition for AI Advantage," commentary, RAND Corporation, March 13, 2025. As of August 6, 2025:
https://www.rand.org/pubs/commentary/2025/03/seeking-stability-in-the-competition-for-ai-advantage.html

Renshaw, Jarrett, and Trevor Hunnicutt, "Biden, Xi Agree That Humans, Not AI, Should Control Nuclear Arms," Reuters, November 16, 2024.

Sabin, Sam, "Exclusive: Anthropic, Feds Test Whether AI Will Share Sensitive Nuclear Info," Axios, November 14, 2024.

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

Scharre, Paul, "Debunking the AI Arms Race Theory," *Texas National Security Review*, Vol. 4, No. 3, Summer 2021.

Schelling, Thomas C., *Arms and Influence*, Yale University Press, 1966.

Scher, Aaron, and Lisa Thiergart, *Mechanisms to Verify International Agreements About AI Development*, Machine Intelligence Research Institute Technical Governance Team, November 27, 2024.

Shaikh, Shaan, and Wes Rumbaugh, "The Air and Missile War in Nagorno-Karabakh: Lessons for the Future of Strike and Defense," Center for Strategic and International Studies, December 8, 2020.

Shavit, Yonadav, "What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring," arXiv, arXiv: 2303:11341, May 30, 2023.

Shoker, Sarah, Andrew Reddie, Sarah Barrington, Ruby Booth, Miles Brundage, Husanjot Chahal, Michael Depp, Bill Drexel, Ritwik Gupta, Marina Favaro, et al., "Confidence-Building Measures for Artificial Intelligence: Workshop Proceedings," arXiv, arXiv:2308.00862, August 3, 2023.

Snyder, Glenn H., *Deterrence and Defense: Toward a Theory of National Security*, Princeton University Press, 1961.

Sutton, Rich, "The Bitter Lesson," webpage, Incompleteideas.net, March 13, 2019. As of August 6, 2025: http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Trager, Robert F., Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, Ranjit Lall, Owen Larter, Seán Ó hÉigeartaigh, et al., *International Governance of Civilian AI: A Jurisdictional Certification Approach*, Oxford Martin Centre for the Governance of AI, 2023.

U.S. Department of Justice, "Chinese National Residing in California Arrested for Theft of Artificial Intelligence-Related Trade Secrets from Google," press release, last updated February 6, 2025.

Vaynman, Jane, and Tristan A. Volpe, "Dual Use Deception: How Technology Shapes Cooperation in International Relations," *International Organization*, Vol. 77, No. 3, Summer 2023.

Vaynman, Jane, and Tristan A. Volpe, "Duplicity and Disclosure: How Technology Shapes Arming Strategies," workshop, Stanford University, June 6, 2024.

Volpe, Tristan A., "Biotechnology and the Dead Zone for Managing Dual-Use Dilemmas," in Nathan A. Paxton, ed., *Disincentivizing Bioweapons*, Nuclear Threat Initiative, 2024.

Wheeler, Nicole E., "Responsible AI in Biotechnology: Balancing Discovery, Innovation and Biosecurity Risks," *Frontiers in Bioengineering and Biotechnology*, Vol. 13, February 4, 2025.

White House, *Winning the Race: America's AI Action Plan*, July 2025.

Whittlestone, Jess, and Jack Clark, "Why and How Governments Should Monitor AI Development," arXiv, arXiv:2108.12427, August 31, 2021.

About the Authors

Jim Mitre is the inaugural vice president and director of RAND Global and Emerging Risks, leading RAND's efforts to deliver public policy research on the most-consequential challenges to civilization and global security. He holds a J.D.

Michael C. Horowitz is a senior political scientist at RAND, the director of Perry World House at the University of Pennsylvania, and the Richard Perry Professor at the University of Pennsylvania. His research focuses on the intersection of AI and other emerging technologies with global politics, military innovation, leadership in international affairs, and methods for geopolitical forecasting. He received a Ph.D. in government.

Natalia Henry is a Ph.D. student in political science at the University of Pennsylvania. Her research focuses on military power, battlefield effectiveness, and the consequences of military technology and innovation. She received her B.A. in history.

Emma Borden is the associate director of the RAND Geopolitics of AGI Initiative and a senior policy analyst. Her previous experience includes work on Europe, Middle East, and homeland defense policy issues. She received her M.S. in foreign service.

Joel B. Predd is a senior engineer at RAND. His current research portfolio focuses on the geopolitics of AI and on the intersection of security and economic competition with China. He holds a Ph.D. in electrical engineering.

Sarah Kreps is the John L. Wetherill Professor in the Cornell University Department of Government, adjunct professor of law at the Cornell Law School, and the director of the Tech Policy Institute in the Cornell Brooks School of Public Policy. Her research focuses on the intersection of technology, international politics, and national security. She received her Ph.D. with a concentration in international relations and security studies.

Miles Brundage is the executive director of the AI Verification and Evaluation Research Institute, a nonresident senior fellow at the Institute for Progress, a member of the AI Policy and Governance Working Group, and a member of the AI Governance Forum at the Center for a New American Security. He holds a Ph.D. in human and social dimensions of science and technology.

James D. Fearon is the Theodore and Frances Geballe Professor in the School of Humanities and Sciences and a professor of political science at Stanford University, as well as senior fellow at the Freeman Spogli Institute for International Studies. His research examines civil and interstate war, deterrence theory, ethnic violence, political factors shaping economic development, and how democratic systems hold leaders accountable. He received his Ph.D. in political science.

Karl P. Mueller is a senior political scientist at RAND and faculty member in the RAND School of Public Policy. He specializes in research related to military and national security strategy, and his recent focus includes the strategic implications of advanced AI. He received his Ph.D. in politics.

Jane Vaynman is an assistant professor of strategic studies at the School of Advanced International Studies at Johns Hopkins University. Her research explores how rival states engage in security cooperation, how arms control agreements are structured, and how technological change influences international collaboration and competition. She received her Ph.D. in political science.

Tristan A. Volpe is an associate professor at the Naval Postgraduate School and a nonresident fellow in the Nuclear Policy Program at the Carnegie Endowment for International Peace. His research examines how technology affects coercion, cooperation, and competition among states. He received a Ph.D. in political science. The views here are the author's own and do not reflect those of the U.S. government.