Artificial Intelligence and the Future of Strategic Stability

Michael C. Horowitz

Texas National Security Review 9, no. 2 (Spring 2026) | © 2026 by Michael C. Horowitz | DOI: 10.1353/tns.00034

How will advances in artificial intelligence impact strategic stability? A growing number of studies and reports assessing the ways that advances in AI could influence global politics focus on the potential risks to strategic stability from integration of AI into the nuclear domain, particularly in large language models and frontier AI. These risks come from multiple potential sources, including miscalculation by machines, sidestepping of human firebreaks to escalation, AI-induced accidents, the speed of AI-enabled warfare, and other mechanisms. The relationship between AI integration and strategic stability may change over time as knowledge and experience with AI systems increases, thus decreasing the likelihood of automation bias, but fundamentally the relationship will depend on second-strike capabilities. While there is inherent uncertainty, since we are still early in the age of AI, at this point it appears as though the higher the confidence nuclear-armed states have in their second-strike capabilities, the lower the probability that they integrate AI in dangerous ways that make escalation at machine speed more likely, and vice versa.

How will advances in artificial intelligence (AI) impact strategic stability? Over the last five years, a growing number of studies and reports have assessed the ways that advances in AI could influence global politics, including the potential risks to strategic stability from integration of AI into the nuclear domain, particularly in large language models (LLMs) and frontier AI.[1] These risks come from multiple potential sources, including miscalculation by machines, sidestepping of human firebreaks to escalation, AI-induced accidents, the speed of AI-enabled warfare, and other mechanisms.[2] These arguments generally make strong negative assumptions about the ways in which uncertainty will influence the development of capabilities and doctrine.

But there are other possible outcomes. Given a different set of assumptions, one can plausibly argue

1 Herbert Lin, "Artificial Intelligence and Nuclear Weapons: A Commonsense Approach to Understanding Costs and Benefits," *Texas National Security Review* 8, no. 3 (2025): 98–109; Edward Geist, *Deterrence Under Uncertainty: Artificial Intelligence and Nuclear Warfare* (Oxford University Press, 2023); Vladislav Chernavskikh and Jules Palayer, "Impact of Military Artificial Intelligence on Nuclear Escalation Risk," *SIPRI*, June 2025, https://www.sipri.org/publications/2025/sipri-insights-peace-and-security/impact-military-artificial-intelligence-nuclear-escalation-risk; Vladislav Chernavskikh, "Nuclear Weapons and Artificial Intelligence: Technological Promises and Practical Realities," *SIPRI*, September 2024, https://doi.org/10.55163/VBQX6088; The Global Commission on Responsible Artificial Intelligence in the Military Domain, *Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of Artificial Intelligence in the Military Domain*, 2025, https://hcss.nl/news/new-gc-reaim-strategic-guidance-report-on-responsible-ai-in-the-military-domain/; Jacob Stokes, Colin H. Kahl, Andrea Kendall-Taylor, and Nicholas Lokker, "Averting AI Armageddon: US-China-Russia Rivalry at the Nexus of Nuclear Weapons and Artificial Intelligence," *Center for a New American Security*, February 13, 2025, https://www.cnas.org/publications/reports/averting-ai-armageddon?; Vincent Boulanin et al., "Artificial Intelligence, Strategic Stability and Nuclear Risk," *SIPRI*, 2020; Michael C. Horowitz, Paul Scharre, and Alexander Velez-Green, "A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence," *arXiv*, 2019, https://doi.org/10.48550/ARXIV.1912.05291; Forrest E. Morgan et al., *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World* (RAND Corporation, 2020), https://doi.org/10.7249/RR3139-1; Edward Geist and Andrew Lohn, *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (RAND Corporation, 2018), https://doi.org/10.7249/PE296; Michael C. Horowitz, "When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability," *Journal of Strategic Studies* 42, no. 6 (September 19, 2019): 764–88, https://doi.org/10.1080/01402390.2019.1621174; Matthijs M. Maas, "How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons," *Contemporary Security Policy* 40, no. 3 (July 3, 2019): 285–311, https://doi.org/10.1080/13523260.2019.1576464.

2 Paul Slovic and Herbert S. Lin, "The Caveman and the Bomb in the Digital Age," in *Effects of the Global Information Ecosystem on the Risk of Nuclear Conflict*, ed. Harold A. Trinkunas, Herbert S. Lin, and Ben Loehrke (Hoover Institution Press, 2020).

that the integration of AI will not, in fact, increase the risk of nuclear escalation and undermine strategic stability. For example, one might argue instead that the increased risks potentially associated with AI could lead to more caution in how decision-makers decide to utilize algorithms or humans in areas like early warning, strike, and more.

Applications of technologies often go through a life cycle involving differences between perceived and actual effectiveness.

How do we reconcile these different possibilities, given that we have only nascent evidence, at best, about the way that militaries will seek to integrate advances in AI into their nuclear systems?[3] This article focuses on the intersection of time, uncertainty, and confidence in capabilities and how it means that *both perspectives* may be correct at different points in the life cycle of particular AI uses.

Applications of technologies often go through a life cycle involving differences between perceived and actual effectiveness. Early in the life cycle, "hype periods" occur wherein perceived effectiveness is greater than actual effectiveness, consistent with arguments from Jon Lindsay and Josh Rovner about the gaps between expectations about technology and the limits of implementation.[4] The relationship between perception and reality can drive mistakes due to automation bias stemming from overconfidence, as users perceive technology as a silver bullet and become less likely to look for and catch errors. If the machine is a calculator, and the activity is splitting a bill at a restaurant, the cost of that overconfidence is low. But if the machine is an algorithm designed to identify missile launches but is trained on insufficient data and at risk of hacking, the potential costs are high. As technology improves over time, however, the relationship can reverse, creating periods of "trust gaps" where the technology is actually more effective than it is perceived to be. The goal is to get to calibration, where there is alignment between expectations about the technology and the reality of the technology.

This life cycle suggests that the relationship between AI and strategic stability in the nuclear domain is not strictly linear. Improved capabilities might increase the safety and reliability of systems in some cases, while in other cases leading to overconfidence that could increase the risk of nuclear accidents and reduce strategic stability. Militaries are often conservative about the integration of new capabilities, but states' choices about AI integration will fundamentally depend on the existing level of confidence that they have in their second-strike capabilities. The higher that level of confidence, the lower the probability, all else being equal, that countries should take dangerous risks with the use of AI in the nuclear domain. Assumptions about the probability of accidents, regime type, and technological progress could also shape the impact of AI on strategic stability.

Psychology and organizational behavior, more than questions of technology itself, will shape how humans engage with AI applications, along with the impact this interaction will have on nuclear risk. These dynamics also illustrate the potential for confidence-building measures (CBMs) to increase the probability that military AI applications increase, rather than decrease, international security. CBMs were tools designed during the Cold War to encourage cooperation between partners and adversaries with shared interests in avoiding accidental and inadvertent war, thus reducing the risk of strategic instability. While not a panacea, such measures as dialogue, norm building, and potential agreements on human involvement with early-warning systems could represent an area where even the United States and China have shared interests: avoiding inadvertent war, as the Biden-Xi agreement in November 2024 on ensuring human control over nuclear weapons shows.[5]

Existing CBMs, like the Political Declaration on Responsible Military Use of AI and Autonomy (endorsed by sixty countries), have started the process

3 For example, France, the People's Republic of China, the United Kingdom, and the United States have all stated that they will not integrate AI into key decision-making involving nuclear weapons. See "Principles and Responsible Practices for Nuclear-Weapon States: Working Paper Submitted by France, the United Kingdom of Great Britain and Northern Ireland and the United States of America," 2020 Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons, July 2022, https://docs.un.org/en/NPT/CONF.2020/WP.70; Jarrett Renshaw and Trevor Hunnicutt, "Biden, Xi Agree That Humans, Not AI, Should Control Nuclear Arms," *Reuters*, November 17, 2024, https://www.reuters.com/world/biden-xi-agreed-that-humans-not-ai-should-control-nuclear-weapons-white-house-2024-11-16/.

4 "Gartner Hype Cycle Research Methodology," Gartner, https://www.gartner.com/en/research/methodologies/gartner-hype-cycle; "Understanding Gartner's Hype Cycles," Gartner, https://www.gartner.com/en/documents/396330; Joshua Rovner, "Strategy and Grand Strategy in New Domains," in *The New Makers of Modern Strategy: From the Ancient World to the Digital Age*, ed. Hal Brands (Princeton University Press, 2023), 1067–91; Jon R. Lindsay, "War Is from Mars, AI Is from Venus: Rediscovering the Institutional Context of Military Automation," *Texas National Security Review* 7, no. 1 (2023): 29–47.

5 Renshaw and Hunnicutt, "Biden, Xi Agreed." On the potential role of cooperation in conflict, see Thomas C. Schelling, *The Strategy of Conflict* (Harvard University Press, 1980), 20.

of creating shared norms for responsible behavior.[6] Another example could be an Autonomous Incidents Agreement, modeled after the Incidents at Sea Agreement, designed to create a forum for information sharing about the deployment of military systems with machine-learning attributes. Such an agreement, which could build on ongoing US-China dialogue on maritime incidents, as well as the original Incidents at Sea Agreement, could contribute to decreasing the risk that accidents involving AI-enabled systems lead to escalation.

In what follows, I first lay out background concerning AI and strategic stability. I discuss the existing arguments in the literature for why advances in AI could threaten strategic stability and the arguments for why that may not be the case. I then introduce a new theoretical framework for how the relationship between technological progress and beliefs about technological progress, over time, can impact the risks that countries are willing to take in integrating AI into aspects of the nuclear domain. I then discuss other key factors that will influence the impact of AI on strategic stability, as well as possibilities for risk reduction.

Background

Scholars and dreamers alike have imagined the idea of machines that could do the work of humans for centuries.[7] But the modern dream of artificial intelligence dates from the end of World War II, when scholars such as Vannevar Bush and Alan Turing proposed the notion of intelligent machines and John McCarthy hosted a 1956 conference that allegedly coined the term "artificial intelligence."[8] Since the 1950s, there have been waves of interest in intelligent machines, with ambition and excitement about advances in computing technology generally yielding to disappointment and what are described as "winters" in AI research.[9]

Current interest in AI stems from advances in research techniques often described as machine learning (ML). As opposed to rule-based AI systems, which derive behaviors and outputs from complicated if/then matrices, AI models assess patterns in data to develop potential solutions to specific problems.

These algorithms are already used widely, from delivering internet search results to predicting trends based on historical data. There are many variants in how ML algorithms work, from supervised algorithms where human coders label training data, to unsupervised algorithms that develop patterns in the data themselves, to deep learning models that use neural networks.[10]

For the purposes of this chapter, "artificial intelligence" refers to tasks conducted by machines that people used to believe required human intelligence. One challenge is that the definition of artificial intelligence—and especially more advanced concepts such as frontier AI or artificial general intelligence (AGI)—is a moving target, almost always referring to what computers cannot yet achieve. For example, scholars once thought that computers that could defeat humans at checkers or chess would be "smart" computers. Once it happened, that became programming. Then scholars thought computers that could win at games such as "Go" would be smart computers. Once that happened, programs like AlphaGo and AlphaGo Zero from DeepMind were also described as computer programs.[11] Thus, it makes sense to be as specific as possible because AI applications may be redefined as computing once they are finalized, with "AI" always defined as what is over the horizon—a phenomenon known as the AI effect.[12]

Furthermore, this chapter focuses on AI use cases—algorithms designed to complete a specific task—rather than AGI, or AI with the capacity to follow broad, nonspecific instructions and design its own subsidiary algorithms to solve new problems.

Turning from AI to strategic stability, the question of what constitutes "strategic stability" remains much clearer in the abstract than in the specific.[13] The essence of strategic stability is the idea of

6 United States Department of State, Bureau of Arms Control, Deterrence, and Stability, "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," February 16, 2023, https://www.state.gov/bureau-of-arms-control-deterrence-and-stability/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy.

7 Bruce G. Buchanan, "A (Very) Brief History of Artificial Intelligence," *Ai/ML Magazine* 26, no. 4 (2005): 53–60.

8 John McCarthy et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955," *AI Magazine* 27, no. 4 (Winter 2026), https://doi.org/10.1609/aimag.v27i4.1904.

9 Luciano Floridi, "AI and Its New Winter: From Myths to Realities," *Philosophy & Technology* 33, no. 1 (March 1, 2020): 1–3, https://doi.org/10.1007/s13347-020-00396-6.

10 Stuart J. Russell, Peter Norvig, and Ernest Davis, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Prentice Hall, 2010).

11 "AlphaGo," https://www.deepmind.com/research/highlighted-research/alphago; Cade Metz, "In Two Moves, AlphaGo and Lee Sedol Redefined the Future," *Wired*, https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/.

12 Michael C. Horowitz and Lauren Kahn, "The Cost of the AGI Delusion," *Foreign Affairs*, September 26, 2025, https://www.foreignaffairs.com/united-states/cost-delusion-artificial-general-intelligence.

13 Thomas C. Schelling, "Surprise Attack and Disarmament," *Bulletin of the Atomic Scientists* 15, no. 10 (December 1, 1959): 413–18, https://doi.org/10.1080/00963402.1959.11454030; Albert Wohlstetter, "The Delicate Balance of Terror," *Foreign Affairs*, 1958, https://www.foreignaffairs.com/articles/1959-01-01/delicate-balance-terror.

first-strike stability to avoid nuclear war, meaning that, in equilibrium, potential adversaries do not feel pressure to launch a preventive strike. While invented to discuss the nuclear balance and pressures surrounding the initiation of nuclear war, in principle, the notion of strategic stability has broader applicability. In a strategically stable situation—even in the midst of a crisis—if both sides wish to avoid war, neither will feel pressure to launch a preemptive strike.[14] Strategic stability does not rule out the possibility of war, but it means that war should only occur if it is deliberate—that is, one or both sides believe that war is the only recourse due to a challenge to their core interests. For the purposes of this article, strategic stability refers to the idea that a set of two (or more) countries perceive a stable balance of power (including capabilities and intentions), meaning there is no need to start or escalate a war unless it is deliberate.

AI and Strategic Stability: The Current Debate

A challenge in assessing the impact of advances in artificial intelligence (especially ML methods) on strategic stability is that we lack the evidence to make an empirical assessment. The integration of AI into the nuclear domain in particular is simply at too early a stage. Nevertheless, researchers have laid out several underlying arguments for why AI may undermine strategic stability, and why it may not.

Why AI May Undermine Strategic Stability

While many aspects of the nuclear weapons infrastructure are already automated, and have been since the 1960s, researchers offer several arguments for why advances in AI could undermine strategic stability.[15] These arguments fall into different areas, including early warning, surveillance, uncrewed vehicles with nuclear weapons, and the impact on strategic stability from conventional military uses of AI.

First, algorithms, though not using machine learning, have already been part of nuclear systems. The 1983 false warning emitted by the Soviet Oko system is the most prominent public example of a machine-related error nearly triggering nuclear war. In that case, the Soviet early-warning system showed an incoming US nuclear attack. Watch officer Lt. Col. Stanislov Petrov, however, rather than reporting a US nuclear attack to his superiors, reported a systems malfunction. Petrov was correct—it was a machine error—but if he had trusted the machine, rather than his own judgment, nuclear war could have easily resulted.[16] That example had nothing to do with ML, but it illustrates the inherent risks of integrating computing.

The Soviet Perimeter system, deployed during the Cold War, also illustrates the logic and risk of machine-based decision-making about nuclear attacks.[17] There are disagreements about how the system functioned and whether it was ever fully deployed.[18] Most accounts, however, suggest that if the system detected a nuclear attack, it would notify the Soviet General Staff and look for a response. If a response was not received, the Perimeter system would transfer nuclear launch authority to local commanders and fire communication rockets to notify watch commanders of the order to launch a nuclear strike at the United States.[19] The Emergency Action Message system in the United States had to be activated by a human, but after activation, it similarly automated the order to launch a nuclear attack and delivered the order to a dispersed commander in the field.[20]

These incidents and systems did not involve AI. AI approaches such as ML involve training algorithms to assess a given situation (like recognizing a particular type of image) based on some set of training data that may or may not be labeled. Whether the

14 Eldridge A. Colby, "Defining Strategic Stability: Reconciling Stability and Deterrence," in *Strategic Stability: Contending Interpretations*, ed. Elbridge A. Colby and Michael S. Gerson (Army War College Press, 2013), 47–84; Elbridge A. Colby and Michael S. Gerson, eds., *Reclaiming Strategic Stability* (Army War College Press, 2013); James M. Acton, "Reclaiming Strategic Stability," in Colby and Gerson, *Strategic Stability*, 117–46.

15 While artificial intelligence as a concept and research area has been around for decades, the current debate stems from the way advances in machine learning, neural networks, and other algorithmic methods are reshaping the ability of machines to process information, and the speed of that processing.

16 Horowitz, Scharre, and Velez-Green, "A Stable Nuclear Future?"; Pavel Aksenov, "Stanislav Petrov: The Man Who May Have Saved the World," *BBC News*, September 26, 2013, https://www.bbc.com/news/world-europe-24280831.

17 Bruce G. Blair, "The Logic of Accidental Nuclear War" (Brookings Institution, 1993); Nicholas Thompson, "Inside the Apocalyptic Soviet Doomsday Machine," *Wired*, September 21, 2009, https://www.wired.com/2009/09/mf-deadhand/.

18 Horowitz, Scharre, and Velez-Green, "A Stable Nuclear Future?"

19 Interview with Gen.-Col. (ret.) Andrian A. Danilevich, March 5, 1990, http://nsarchive.gwu.edu/nukevault/ebb285/vol%20iI%20Danilevich.pdf, 62–63; interview with Viktor M. Surikov, September 11, 1993, http://nsarchive.gwu.edu/nukevault/ebb285/vol%20II%20Surikov.pdf, 134–35. See also Michael T. Klare, "'Skynet' Revisited: The Dangerous Allure of Nuclear Command Automation," Arms Control Association, April 2020, https://www.armscontrol.org/act/2020-04/features/skynet-revisited-dangerous-allure-nuclear-command-automation; Thompson, "Inside the Apocalyptic Soviet Doomsday Machine."

20 Blair, "The Logic of Accidental Nuclear War."

task is identifying a cat or a missile launch, an AI algorithm will use its training to complete the task, often reporting the level of confidence in its approach. For example, in an attempt to get faster warning of a nuclear attack, AI approaches could be used for pattern recognition to detect missile launches. Using an AI approach presents risks, however. Any algorithms would have to be trained on simulated data (whether supervised or unsupervised methods are used), since there is not a huge dataset of nuclear attacks to draw from, which increases the risk of an accident or other flaw in the algorithm itself. Early-warning algorithms might also be vulnerable to hacking or spoofing by adversaries or third parties. Critically, early-warning algorithms based on AI approaches, if trusted completely, might lead countries to decrease the number of watch officers or cut them out of the loop entirely, meaning there would be no Petrov to prevent the next early-warning accident from escalating. Or, alternatively, as described in more detail below, automation bias due to overconfidence in an AI system could cause a future Petrov to outsource cognitive judgment to the machine and thus make them less likely to assume that the machine made an error, as Petrov did in 1983.[21]

A second way in which AI could decrease strategic stability and increase nuclear risk is if the enhanced surveillance capacity of artificial intelligence placed pressure on second-strike capabilities, in combination with advances in the speed and accuracy of conventional munitions. The combination of uncrewed aerial vehicles and autonomous systems could allow for distributed, real-time tracking of adversary assets on land, for example. Drones offer the potential for persistent surveillance of mobile transporter erector launchers (TELs), helping enable targeting of TELs and increasing the potential effectiveness of counterforce strikes.[22] Algorithms could process and integrate intelligence, surveillance, and reconnaissance (ISR) information across multiple data sources quickly enough to enable near-real-time tracking.

Meanwhile, at sea, distributed pre-placed sensors in chokepoints such as the GIUK gap (encompassing Greenland, Iceland, and the United Kingdom) could provide new opportunities for actors to track adversary submarines with nuclear weapons (SSBNs) by providing them with early or real-time warning of SSBN transits.[23] While tracking adversary submarines might seem tactically attractive, the macro effect could be to undermine national confidence in second-strike capabilities, increasing instability. In addition, AI pattern and image recognition could improve anti-submarine warfare (ASW) activities more broadly, from modeling where to do searches, to rapidly reacting, to improving tracking. For example, Task Force Ocean, run by the US Office of Naval Research, sought to build oceanographic datasets and apply reinforcement learning to outperform humans at tracking tasks.[24]

The combination of uncrewed aerial vehicles and autonomous systems could allow for distributed, real-time tracking of adversary assets on land, for example.

Whether on land or at sea, undermining national confidence in second-strike capabilities such as SSBNs and mobile TELs through sensors enabled and linked together with machine learning could increase the risk that nuclear powers feel the pressure to strike first in order to avoid adversaries using AI-based tracking methods that target their second-strike capabilities in a crisis.

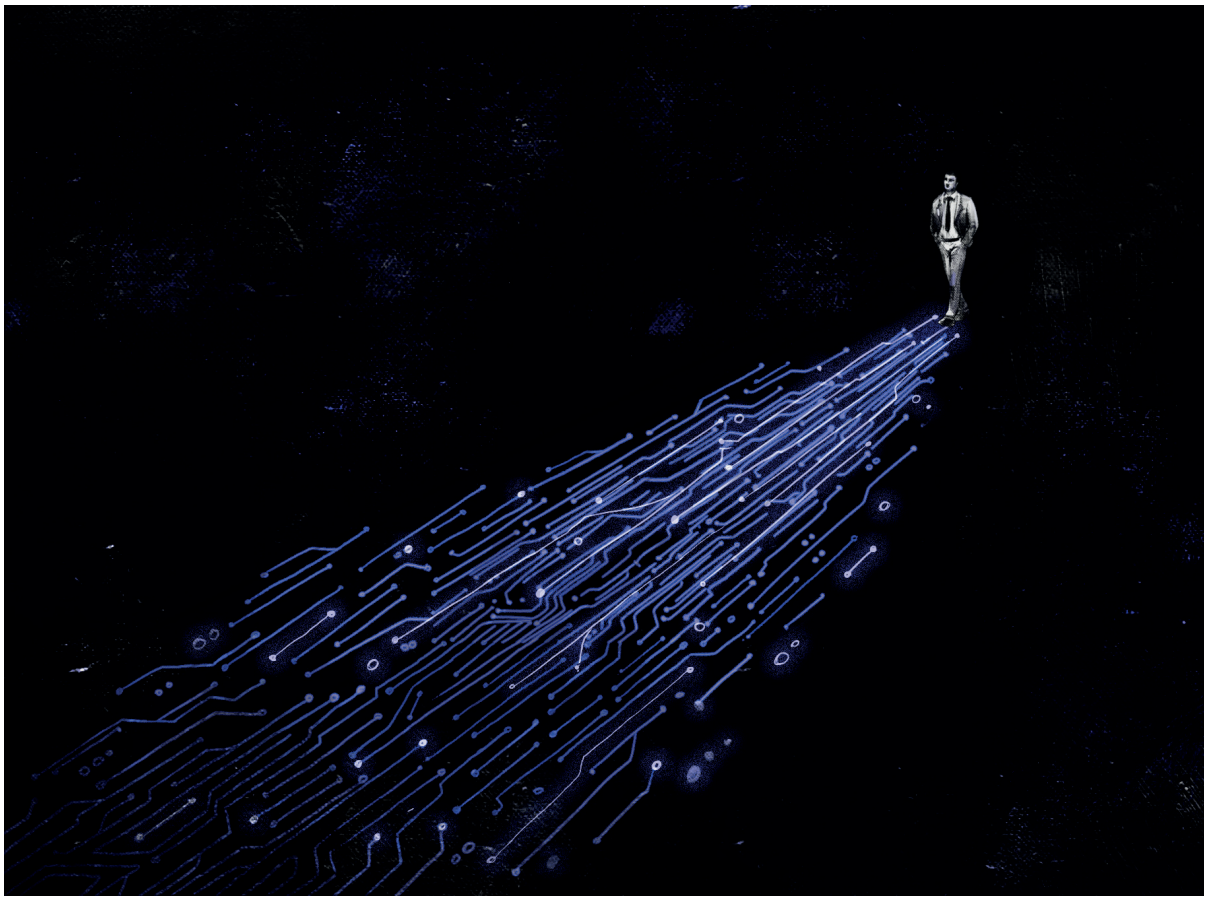
A third pathway through which advances in AI could undermine strategic stability is if they give

21 Linda J. Skitka, Kathleen Mosier, and Mark D. Burdick, "Accountability and Automation Bias," *International Journal of Human-Computer Studies* 52, no. 4 (April 1, 2000): 701–17, https://doi.org/10.1006/ijhc.1999.0349; Linda J. Skitka, Kathleen Mosier, and Mark D. Burdick, "Does Automation Bias Decision-Making?," *International Journal of Human-Computer Studies* 51, no. 5 (November 1, 1999): 991–1006, https://doi.org/10.1006/ijhc.1999.0252; Kathleen L. Mosier et al., "Automation Bias: Decision Making and Performance in High-Tech Cockpits," *The International Journal of Aviation Psychology* 8, no. 1 (January 1, 1998): 47–63, https://doi.org/10.1207/s15327108ijap0801_3; Mary Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems," in *AIAA 1st Intelligent Systems Technical Conference* (American Institute of Aeronautics and Astronautics, 2004), https://doi.org/10.2514/6.2004-6313; Mary (Missy) Cummings, "Informing Autonomous System Design Through the Lens of Skill-, Rule-, and Knowledge-Based Behaviors," *Journal of Cognitive Engineering and Decision Making* 12, no. 1 (2018): 58–61, https://doi.org/10.1177/1555343417736461; David Lyell et al., "Automation Bias in Electronic Prescribing," *BMC Medical Informatics and Decision Making* 17, no. 1 (March 16, 2017): 28, https://doi.org/10.1186/s12911-017-0425-5; Horowitz, Scharre, and Velez-Green, "A Stable Nuclear Future?"

22 Keir A. Lieber and Daryl G. Press, "The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence," *International Security* 41, no. 4 (April 2017): 9–49, https://doi.org/10.1162/ISEC_a_00273.

23 Andrew Metrick, "(Un)Mind the Gap," *US Naval Institute* 145, no. 10 (October 1, 2019), https://www.usni.org/magazines/proceedings/2019/october/unmind-gap; Jonathan Gates, "Is the SSBN Deterrent Vulnerable to Autonomous Drones?," *The RUSI Journal* 161, no. 6 (November 1, 2016): 28–35, https://doi.org/10.1080/03071847.2016.1265834.

24 Patrick Tucker, "How AI Will Transform Anti-Submarine Warfare," *Defense One*, July 2019, https://www.defenseone.com/technology/2019/07/how-ai-will-transform-anti-submarine-warfare/158121/.



leaders such confidence that they deploy uncrewed platforms armed with nuclear weapons. This scenario could eliminate positive human control over nuclear use and increase the risk of accidents. In 2018, Russian President Vladimir Putin spoke about a Russian system under development that would be an "unmanned underwater vehicle . . . that would carry massive nuclear ordnance." More a torpedo than a platform, the system is named Poseidon (formerly called Status-6); in theory, it could be deployed for years near the coast of the United States or another potential adversary, creating a fast and reliable means of nuclear attack for Russia. There is controversy over whether this system exists, and if it does, how close it is to deployment. Russian sources have also speculated about developing and deploying uncrewed bombers with nuclear weapons,[25] an option the United States decided not to pursue with its B-21 program. In 2016, while the B-21 was in development, Robin Rand, then-head of Air Force Global Strike Command, stated: "We're planning on [the B-21] being manned. . . . I like the man in the loop . . . very much, particularly as we do the dual-capable mission with the nuclear weapons."[26]

Uncrewed platforms with nuclear weapons could offer advantages that make them attractive to states. The endurance offered by AI-enabled bombers or submarines could aid in persistent deployments that states may believe enhance their capabilities. If bombers, through air-to-air refueling, could stay in the air for weeks or months, or submarines could stay at sea for years, because there are no humans on board, countries might initially feel more secure. The fear is that, without humans on board, these platforms would be at a higher risk for accidents in crisis situations, because there would not be a human directly engaged to make judgment calls. Overconfidence in AI could also contribute to such a mindset through automation bias.

Another way that AI integration could threaten strategic stability is through pressure generated on national command systems by the increasing speed

25 "Russia Could Deploy Unmanned Bomber After 2040—Air Force," https://www.globalsecurity.org/wmd/library/news/russia/2012/russia-120802-rianovosti01.htm.

26 Hope Hodge Seck, "Air Force Wants to Keep 'Man in the Loop' with B-21 Raider," Military.com, September 19, 2016, https://www.military.com/defensetech/2016/09/19/air-force-wants-to-keep-man-in-the-loop-with-b-21-raider; US Air Force, "RPA Vector: Vision and Enabling Concepts, 2013–2038," Headquarters, US Air Force, February 17, 2014, https://www.globalsecurity.org/military/library/policy/usaf/usaf-rpa-vector_vision-enabling-concepts_2013-2038.pdf.

of conventional warfare.[27] Part of the incentive for the adoption of AI for militaries is the ability to operate at "machine speed,"[28] as former Deputy Secretary of Defense Robert Work said—allowing a military to get inside the decision loop of an opponent. Militaries will want to use algorithms for pattern recognition, coordination of human and machine assets, decision aids for commanders, autonomous platforms, and more. In a crisis or a conventional conflict, countries may fear that the increasing pace of warfare means that they could lose so quickly that they will not have time to make a reasoned decision about the use of nuclear weapons, even to avoid decapitation.

Thus, fear of the increasing speed of conventional warfare, driven by AI, could create incentives for less stable nuclear launch postures, such as launch on warning and pre-delegation. Other types of technological change perceived to increase the speed of warfare, such as hypersonics, could have related effects, but what is unique about AI here is the intersection of increasing speed with the potential for overconfidence and the offloading of judgment to an algorithm.[29] Psychologically, speed does not just place pressure on decision-makers; it also makes them more likely to use what psychologist Daniel Kahneman calls System 1 thinking. Kahneman famously describes the rapid judgments of System 1 behavior and contrasts them with the slower, more deliberate System 2 approach.[30]

Why AI May Not Undermine Strategic Stability

By contrast, there are several reasons to think that the connections between advances in AI and nuclear risk are exaggerated, and that AI will decrease, or at least not increase, the risk of deliberate or inadvertent nuclear war. Many of the arguments concerning how applications of AI will increase nuclear danger assume suboptimal behavior on the part of states and individuals. This argument is not to say that states always behave in ways implied by strongly rational theories of state behavior. But some of the arguments concerning how AI could generate nuclear danger presume states not just acting suboptimally, but in ways that seem to contradict other obvious logics that govern military behavior, especially the desire of leaders to avoid tying their hands when it comes to conflict escalation, so that they can control potential nuclear use.[31]

Most important, militaries want their systems to work.

While it certainly makes sense that fully automating early warning would increase the risk of an accident, it raises the question of why militaries would automate early warning in the first place. After all, modern militaries have their systems undergo strict testing and evaluation procedures. Fear that it may be difficult to fully validate systems with integrated AI in critical systems is a key obstacle to the adoption of AI in the American military, for example.[32] Given the consequences of launching a nuclear strike by accident, countries have introduced complex systems to govern the potential use of nuclear weapons and ensure that any use is deliberate. These systems may themselves still fail to decrease, or may even increase, the probability of an accident,[33] but the reason is not because militaries are not paying attention, at least in this case. The US Department of Defense, for example, explicitly declared in the *2022 Nuclear Posture Review* that "the United States will maintain a 'human in the loop' for all actions critical to informing and executing decisions by the President to initiate and terminate nuclear weapons employment."[34]

Most important, militaries want their systems to work. Systems with large-scale safety flaws and vulnerabilities, by definition, do not work well. One might counter that cyber vulnerabilities are well known, yet still exist in the information systems of militaries around the world. But there is a distinction between a cyber system that works well unless hacked

27 Horowitz, "When Speed Kills."

28 Robert O. Work, "Remarks by Defense Deputy Secretary Robert Work at the CNAS Inaugural National Security Forum," Center for New American Security, December 14, 2015, https://www.cnas.org/publications/transcript/remarks-by-defense-deputy-secretary-robert-work-at-the-cnas-inaugural-national-security-forum.

29 Perception here, rather than reality, is the key to influencing national behavior.

30 Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011); Keren Yarhi-Milo, *Knowing the Adversary: Leaders, Intelligence, and Assessment of Intentions in International Relations* (Princeton University Press, 2014).

31 Dan Reiter, *Untied Hands: How States Avoid the Wrong Wars* (Cambridge University Press, 2025).

32 Michele Flournoy, Avril Haines, and Gabrielle Chefitz, *Building Trust Through Testing* (Center for Security and Emerging Technology, Georgetown University, 2020).

33 Scott D. Sagan, *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons* (Princeton University Press, 1995); Blair, "The Logic of Accidental Nuclear War"; Charles Perrow, *Normal Accidents: Living with High-Risk Technologies* (Princeton University Press, 1999).

34 Department of Defense, *2022 Nuclear Posture Review*, 2022, 13, https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.pdf.

and an algorithm that might fail due to inadequate training data or attempts to apply the algorithm inappropriately, in addition to the risk of hacking. Thus, while militaries that truly feel at risk of decapitation might be more likely—in theory—to consider fully automating their early-warning systems, that outcome seems less likely in reality, as militaries are increasingly aware of the safety issues and vulnerabilities associated with algorithms.[35] Moreover, militaries like to have full control over their weapons systems, particularly those with the ability to cause mass destruction. Thus, they have strong incentives to avoid delegating full control to machines and to train human early-warning operators to avoid automation bias.[36] Additionally, warfare is rife with human-induced errors. Tired, angry, or undertrained humans make mistakes that cost lives. The integration of algorithms might reduce reliance on humans in ways that actually improve performance.[37]

If it is possible to develop effective algorithms for early warning, algorithms could also have benefits in terms of the speed of pattern recognition. Especially as the pace of warfare grows, militaries worry about the decreased time between launch and impact, driven by faster missiles. Faster pattern recognition of a launch could buy back time for the decision-maker to consider how to respond. Essentially, speed of detection could ease the cognitive load, allowing more System 2 thinking to occur during a crisis.[38] Buying time could reduce pressure, in theory, for a hasty decision to launch a preemptive strike in response to a false warning.

Will AI-driven surveillance undermine second-strike capabilities on land or at sea? It is unclear. For mobile TELs or SSBN tracking, issues unrelated to AI will make counterforce strikes difficult. For mobile TELs, executing a strike is not as simple as knowing where a TEL is at time *T*. The launching of a massive strike on adversary TELs would require not only simultaneous tracking of those TELs across a potentially large adversary nation, but also the communication of that information to strike assets. That communication and then the launching of those strike assets to hit the mobile TELs would take time, at which point those mobile TELs would have moved. Now, one could make assumptions about simultaneous real-time tracking and striking, but that thinking places a lot of faith in uncertain technologies.[39] The number of things that could go wrong—particularly when an adversary would presumably be actively trying to hide their assets (excluding a true bolt-from-the-blue scenario)—makes an attack risky enough that an attacker is likely to think twice.

Additionally, it is far from clear that unmanned underwater vehicles (UUVs), combined with sensors, will generate a level of transparency of the oceans sufficient to undermine the reliability of SSBNs. These systems would need to be quiet and persistent to avoid being detected and tracked. The limited size, weight, and power of these vehicles, along with the physics of communication in the undersea environment, will make real-time tracking and fixing of SSBNs difficult.[40] The number of sensors needed to monitor SSBNs, even British and Chinese SSBNs that have to go through known chokepoints, would also be so large as to require advances in battery power, electronics miniaturization, and other related technology, not just AI.[41]

What about uncrewed platforms armed with nuclear weapons? These systems may arguably enhance, rather than detract from, strategic stability. For countries with limited numbers of nuclear platforms and weapons, uncrewed platforms that could stay in the air or at sea for months or longer could offer greater second-strike reliability than systems that must return to bases more frequently. But most countries are unlikely to develop uncrewed systems with nuclear weapons if they think it would decrease their control over the systems—the Russian case aside (admittedly a large caveat). Trust surrounding the systems with the ability to launch nuclear weapons is paramount—much more important than

35 This also suggests that there could be an AI adoption gap at times that leads to tech-infused civil-military relations issues. Especially early in the age of AI, it is possible to imagine conflict between military reticence to adopt unproven technologies and the desire of states to accelerate AI adoption, given expectations of wars fought at machine speed.

36 J. Elin Bahner, Anke-Dorothea Hüper, and Dietrich Manzey, "Misuse of Automated Decision Aids: Complacency, Automation Bias and the Impact of Training Experience," *International Journal of Human-Computer Studies* 66, no. 9 (2008): 688–99; Juergen Sauer, Alain Chavaillaz, and David Wastell, "Experience of Automation Failures in Training: Effects on Trust, Automation Bias, Complacency and Performance," *Ergonomics* 59, no. 6 (June 2, 2016): 767–80, https://doi.org/10.1080/00140139.2015.1094577.

37 Eric Talbot Jensen, "The (Erroneous) Requirement for Human Judgment (and Error) in the Law of Armed Conflict," *SSRN Electronic Journal*, 2020, https://doi.org/10.2139/ssrn.3548314.

38 Kahneman, *Thinking, Fast and Slow.*

39 James N. Miller and Richard Fontaine, "A New Era in US-Russian Strategic Stability: How Changing Geopolitics and Emerging Technologies Are Reshaping Pathways to Crisis and Conflict," CNAS and Harvard Kennedy School, September 2017.

40 Gates, "Is the SSBN Deterrent Vulnerable to Autonomous Drones?"; Horowitz, Scharre, and Velez-Green, "A Stable Nuclear Future?"; Owen R. Cote Jr., "Invisible Nuclear-Armed Submarines, or Transparent Oceans? Are Ballistic Missile Submarines Still the Best Deterrent for the United States?," *Bulletin of the Atomic Scientists* 75, no. 1 (2019): 30–35, https://doi.org/10.1080/00963402.2019.1555998.

41 Horowitz, Scharre, and Velez-Green, "A Stable Nuclear Future?"; John Gower, "Concerning SSBN Vulnerability—Recent Papers," *BASIC*, June 10, 2016, https://basicint.org/blogs-rear-admiral-john-gower-cb-obe-06-2016-concerning-ssbn-vulnerability-recent-papers/.

testing and validation for regular weapons systems—so the issues with algorithms that make trusting them challenging in the best of circumstances will certainly apply to systems with nuclear weapons.

Finally, countries already account for the risk of conventional defeat in war plans involving nuclear weapons. Scenarios that require the use of nuclear weapons often involve things going wrong conventionally, which means countries are already factoring stressful and short-timed situations into their nuclear weapons use contingencies. Even rapid conventional defeat would not necessarily undermine nuclear deterrence. This logic could be complicated, though, by comingled nuclear and conventional command and control systems, which could place nuclear systems at risk even in a purely conventional conflict.

The Role of Trust and Confidence

As the above discussion illustrates, a challenge with the current discourse about AI and strategic stability is that one can make logical arguments on both sides. The question then becomes how to mediate between these claims and make the best possible judgments about what is most likely, especially given that there is not empirical evidence to adjudicate competing claims. These judgments are made more difficult by uncertainty about the likely trajectory of advances in AI itself. Even elite AI researchers disagree about how quickly advances will occur and the types of advances that are possible.[42]

Moreover, the question of how humans will interact with advances in AI, and how that influences nuclear risk, is one of psychology and organizational behavior as much as technology.[43] Cognitive biases fundamentally shape human behavior, particularly the way humans react under stress and challenging circumstances.[44] Especially when these biases interact in an uncertain environment, they can increase the risk of miscalculation. Moreover, organizational routines can interact with biases at the individual level in ways that lead to unpredictable outcomes.[45]

Given the early-stage character of AI as applied in the national security arena in general, it is hard to make a strong empirical claim, or even a weak one, about how advances in algorithms will influence strategic stability in the nuclear domain. Moreover, as explained above, there are theoretical reasons to believe simultaneously that advances in AI will increase nuclear risk and that those risks are overstated. How do we resolve this issue, especially given genuine uncertainty about the trajectory of advances in AI?

Below, I introduce the ideas of time and confidence, adding them to the notion of uncertainty to produce theoretically driven predictions about when the integration of AI in the nuclear domain will decrease strategic stability and when it will not.

Figure 1 presents a three-stage conceptual model focused on the relationship between perceptions of technological effectiveness and reality within a relevant community. These stages are as follows.

- The first stage includes investments surviving the almost inevitable hype cycles that accompany emerging technologies. For example, Robert Fulton pitched the submarine to the British Admiralty in 1815, before the deployment of the first naval vessel powered by steam and ninety-nine years before German deployment of the submarine at the outset of World War I finally caught up to the reality Fulton pictured. In the first period, you also get overconfidence in technologies that have limitations that can lead to suboptimal behavior—like driving your car off a bridge while following Apple Maps directions. This overconfidence results from Clarke's Third Law, which states that "any sufficiently advanced technology is indistinguishable from magic." But electricity goes out, computers crash, and accidents happen. Successful technology adoption requires not simply tossing the old out for the new, but systems integration.
- Second, even when technologies improve, following that overconfidence, there can be a trust gap, as new technologies seem uncertain and unproven, especially if there are accidents that decrease trust. Almost a decade ago, for instance, survey research suggested that soldiers on the ground calling for close air support prefer that support to be delivered by inhabited aircraft rather than drones, even when the platforms have identical capabilities.[46]

42 Katja Grace et al., "When Will AI Exceed Human Performance? Evidence from AI Experts," *arXiv*, 2017, https://doi.org/10.48550/ARXIV.1705.08807.

43 Robert Jervis, "Perception and Misperception in International Politics," in Robert Jervis, *Perception and Misperception in International Politics* (Princeton University Press, 2017). See also Rose McDermott in this volume.

44 Kahneman, *Thinking, Fast and Slow.*

45 Diane Vaughan, "The Dark Side of Organizations: Mistake, Misconduct, and Disaster," *Annual Review of Sociology* 25, no. 1 (August 1, 1999): 271–305, https://doi.org/10.1146/annurev.soc.25.1.271.

46 Julia Macdonald and Jacquelyn Schneider. "Battlefield Responses to New Technologies: Views from the Ground on Unmanned Aircraft," *Security Studies* 28, no. 2 (2019): 216–49.

- Third, increasing knowledge and experience leads to greater calibration, meaning more convergence between expectations about technology and the reality of technology.

er and linear improvement in capabilities, which allows expectations about the technology to rise more gradually with the technology. Nevertheless, the conceptual model remains powerful, influencing

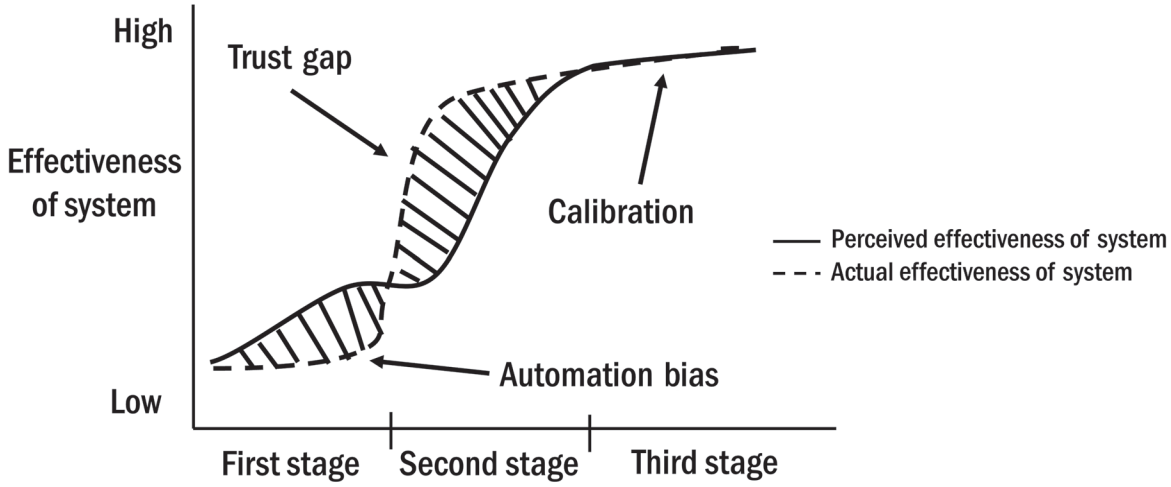


Figure 1. Graph showing the relationship between trust, confidence, and technology adoption. Figure by author.

In short, the three stages can be called the automation bias period, the trust gap period, and the calibrated period.

The first stage reflects the hype period that many identify as characteristic of emerging technologies today, and especially AI.[47] The Gartner hype cycle, a well-known model of technological development in the business world, identifies the way that overenthusiasm about technology, especially before the technology matures, generates a "peak of inflated expectation" that soon yields a "trough of disillusionment." As the technology matures, it then leads to a "plateau of productivity" that is neither as high as early expectations nor as low as the period of disillusionment.[48]

Empirically, not all technologies follow the Gartner hype cycle; examples include MP3 players and DVD players, which grew more steadily in line with technological progress. Technologies that do not go through the hype cycle tend to feature slow-

business leaders and scholars.[49] In the first stage, the risks of negative consequences can be low if there is a failure of the technology to match expectations in a way that reduces actual fielding by militaries.

When fielding occurs, however, the first period can also feature overconfidence, when perceptions of technological effectiveness once again exceed actual effectiveness. Research in international relations clearly demonstrates the way that overconfidence can skew decision-making in ways that are dangerous in international politics, making war and miscalculation more likely.[50] Studies of the business world by Mark Simon and Rodney Shrader show that overconfidence leads to great willingness to pursue breakthrough products in small businesses, while overconfident CEOs are more likely to pursue technological innovation for their firms.[51] However, overconfidence can also lead to more negative risk-taking as well as increasing volatility, and makes the most sense

47 Michael C. Horowitz and Lauren Kahn, "The Cost of the AGI Delusion."

48 "Understanding Gartner's Hype Cycles"; "Gartner Hype Cycle Research Methodology."

49 Ozgur Dedehayir and Martin Steinert, "The Hype Cycle Model: A Review and Future Directions," *Technological Forecasting and Social Change* 108 (2016): 28–41, https://doi.org/10.1016/j.techfore.2016.04.005; Heini Jarvenpaa and Saku J. Makinen, "Empirically Detecting the Hype Cycle with the Life Cycle Indicators: An Exploratory Analysis of Three Technologies," *2008 IEEE International Conference on Industrial Engineering and Engineering Management*, 2008, 12–16, https://doi.org/10.1109/IEEM.2008.4737823.

50 Dominic D. P. Johnson, *Overconfidence and War: The Havoc and Glory of Positive Illusions* (Harvard University Press, 2004); Dominic D. P. Johnson et al., "Overconfidence in Wargames: Experimental Evidence on Expectations, Aggression, Gender and Testosterone," *Proceedings: Biological Sciences* 273, no. 1600 (2006): 2513–20; Dominic D. P. Johnson et al., "Dead Certain," *Human Nature* 23, no. 1 (March 1, 2012): 98–126, https://doi.org/10.1007/s12110-012-9134-z.

51 Mark Simon and Rodney C. Shrader, "Entrepreneurial Actions and Optimistic Overconfidence: The Role of Motivated Reasoning in New Product Introductions," *Journal of Business Venturing* 27, no. 3 (2012): 291–309, https://doi.org/10.1016/j.jbusvent.2011.04.003; Holger Herz, Daniel Schunk, and Christian Zehnder, "How Do Judgmental Overconfidence and Overoptimism Shape Innovative Activity?," *Games and Economic Behavior* 83 (January 1, 2014): 1–23, https://doi.org/10.1016/j.geb.2013.11.001.

in industries characterized by a relatively larger degree of change.[52] Put another way, overconfidence in technological innovation, specifically, decreases the risk of false negatives (since you will err on the side of innovation), but increases the risk of false positives, meaning risky investments.[53] Moreover, these results are not just based on research on US and European firms—studies of Chinese firms also show that overconfidence makes investments in technological innovation more likely, though only in high-tech, non-state-controlled areas.[54]

The impact of overconfidence on willingness to make risky bets on technology has particular importance with respect to AI, as overconfidence in the deployment of algorithms means increasing the risk of automation bias. Automation bias is the phenomenon whereby humans trust machines and algorithms too much, cognitively outsourcing judgments that humans are better placed to make. For example, while machines are better at simultaneous and repetitive tasks, humans are better at improvising and making judgments in situations with incomplete information.[55] Examples of automation bias include the 2009 Air France crash, wherein pilots trusted the autopilot, which said they were not crashing, over their own judgment and training, which correctly suggested the danger.[56] Another example of automation bias comes from the 2003 Patriot missile fratricides, where false confidence in friend or foe identification in the missile-tracking system led to friendly fire, with tragic results.[57]

Research by Linda Skitka, Kathleen Mosier, and Mark Burdick finds that even when automated systems are only advising humans, the advice can lead to error due to human overconfidence. The authors write: "Participants in non-automated settings out-performed their counterparts with a very but not perfectly reliable automated aid on a monitoring task. Participants with an aid made errors of omission (missed events when not explicitly prompted about them by the aid) and commission (did what an automated aid recommended, even when it contradicted their training and other 100% valid and available indicators)."[58]

This first stage is where the probability of negative consequences for strategic stability from uses of AI integration is largest, and where it is most important to try to generate successful calibration that aligns beliefs in the technology with the reality of the technology. Militaries might otherwise be overconfident in trusting algorithms with early-warning tasks, or to believe that AI provides them with surveillance advantages that make them more likely to take chances in crisis situations.

The second period is the trust gap period, during which expectations of technological success have declined, but the technology continues to improve. This stage generates the opposite issue of the first technological development period, because now potential adopters have less trust in the technology than they should, given technological progress.

A historical example of a trust gap comes from the delayed adoption of breech-loading and repeating rifles by the US Union Army during the Civil War. Despite the technological maturity of breech-loading and repeating rifles by the outset of the war, leading Union generals did not trust in the effectiveness of the technology, instead preferring soldiers to use more familiar muzzle-loading technology. Research shows that this delayed adoption likely had a negative impact on Union Army performance during the war.[59]

In the second stage, the highest risk connected to AI and strategic stability involves false negatives, where the natural caution of militaries means that they will trust potential advances in AI less than they should, all else being equal. To the extent that the increased information and warning time from faster AI information processing could buy decision-makers time that leads to detection of false warnings, for example, trust gaps could make detecting those errors less likely, increasing the risk of accidents.

The third period is calibration, where the passage of time, coupled with growth in knowledge and experience with new technologies, means that there

52 David Hirshleifer, Angie Low, and Siew Hong Teoh, "Are Overconfident CEOs Better Innovators?," *The Journal of Finance* 67, no. 4 (August 1, 2012): 1457–98, https://doi.org/10.1111/j.1540-6261.2012.01753.x.

53 Mark Simon and Susan M. Houghton, "The Relationship Between Overconfidence and the Introduction of Risky Products: Evidence from a Field Study," *Academy of Management Journal* 46, no. 2 (2003): 139–49.

54 Wang Shanhui, Wang Zongjun, and Tian Yuan, "The Relationship Between Managerial Overconfidence and Technological Innovation Investment," *Science Research Management* 34, no. 5 (2013): 1–9.

55 Cummings, "Informing Autonomous System Design."

56 Jamie P. Brown, "The Effect of Automation on Human Factors in Aviation," *The Journal of Instrumentation, Automation and Systems* 3, no. 2 (2016): 31–46.

57 John K. Hawley, *Looking Back at 20 Years of MANPRINT on Patriot: Observations and Lessons* (US Army Research Laboratory, 2007).

58 Skitka, Mosier, and Burdick, "Does Automation Bias Decision-Making?"

59 Adam M. Jungdahl and Julia M. Macdonald, "Innovation Inhibitors in War: Overcoming Obstacles in the Pursuit of Military Effectiveness," *Journal of Strategic Studies* 38, no. 4 (June 7, 2015): 467–99, https://doi.org/10.1080/01402390.2014.917628.

is greater alignment between expectations about a technology and the reality of that technology.[60] In the third period, successful adoption and use become more likely, and actors become the least likely to have either automation bias that increases risks through adoption, or trust gaps that increase risk through non-adoption.

To the extent that the increased information and warning time from faster AI information processing could buy decision-makers time that leads to detection of false warnings, for example, trust gaps could make detecting those errors less likely, increasing the risk of accidents.

Of course, in reality, the results could vary for individuals or organizations based on the level of knowledge and experience in working with AI. For example, Lauren Kahn and I find that low levels of AI knowledge and experience lead to a form of the Dunning-Krueger effect, in which there is overconfidence in AI and a higher probability of automation bias.[61] Greater knowledge and experience can therefore mitigate automation bias even at earlier stages, and potentially avoid trust gaps as well.

Additionally, there might be cases where gaps between expectations and reality, though seemingly inefficient, actually promote stability. For example, a small trust gap where people have questions about an algorithm that is actually effective could be a hedge against automation bias when it comes to missile-attack detection.

Another issue is that accidents due to overconfidence might lead to backsliding in AI adoption. After all, the model in figure 1 is notional. In the real world, adoption of a system depends not just on initial experimentation and procurement, but also on how use proceeds over time. One risk with emerging technologies is adoption backsliding, when early accidents and issues that naturally arise in the course of technological development generate bureaucratic opposition to the technology as a whole. This scenario was arguably a key driver of previous AI winters, when investment in AI dried up due to a failure to meet lofty expectations.

The logical next question becomes: What can policymakers do to encourage more alignment between expectations about technological effectiveness and actual effectiveness? One key part of the answer is training and education. The more that AI users—whether early-warning operators, commanders in the field, or senior policymakers receiving advice from decision-support algorithms—understand the capabilities and limitations of algorithms, the better their ability to use algorithms in appropriate ways, regardless of the application.[62] Generational change and more clarity about the range of what is possible surrounding algorithms will help reduce gaps between expectations and reality. The more people who have familiarity with algorithms in their daily lives and work lives, the easier some of those calibration tasks are likely to become. Thus, generational cohort effects could help alleviate some of the challenges above.

Other Key Factors in the Relationship Between AI and Strategic Stability

Other factors influence the intersection of trust and confidence when it comes to AI adoption, and thus how to navigate competing perspectives on the impact on strategic stability. First, what is most likely when it comes to technological development? This baseline question will drive the possible outcomes and some of the choices that states make. Second, what is the level of confidence that countries have in their second-strike capabilities, which will influence their baseline interest in taking risks through using AI? Third, how might different kinds of political regimes, particularly authoritarian regimes, evaluate the benefits and risks of attempting to centralize control through AI algorithms? Finally, how do different assumptions about the baseline risk of nuclear accidents influence how actors should think about AI integration in the nuclear domain?

Progress in AI Research

The first factor is the technology itself. From a relative power perspective, the winners during periods

60 Rovner, "Strategy and Grand Strategy in New Domains."

61 Michael C. Horowitz and Lauren Kahn, "Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts," *International Studies Quarterly* 68, no. 4 (2024), https://doi.org/10.1093/isq/sqae020.

62 Michael C. Horowitz and Lauren Kahn, "The AI Literacy Gap Hobbling American Officialdom," *War on the Rocks*, January 14, 2020, http://warontherocks.com/2020/01/the-ai-literacy-gap-hobbling-american-officaldom/.

of technological change are generally those able to develop strategy, doctrine, and concepts of operation to take advantage of technological change, rather than those that focus purely on technology.[63] Put another way, leadership in technology itself is less than half the battle when it comes to the relationship between technology and military power.[64]

However, when it comes to how AI may influence military power, and thus strategic stability, there are basic technological questions surrounding the range of possible outcomes that will likely determine whether those pessimistic about the impact of AI or those more neutral are correct. Given that AI is a general-purpose technology, as described above, technological trajectories will look different for different applications of AI. And there is large-scale disagreement, even among AI experts, about the timelines for AI to achieve certain milestones. For example, while the average respondent in a 2017 survey of AI researchers, conducted by Katya Grace et al., thought that an algorithm could conduct math research itself in about forty-five years, the distribution of responses ranged from about fifteen years to almost eighty-five years.[65]

The question of technological trajectory matters not just because it defines what is possible, but also because it will influence questions surrounding AI safety. Concerns about a race to the bottom on AI safety, described above, assume that the state of AI is such that countries will feel pressure to deploy algorithms in areas where the technology is not yet ready.[66] If technological development occurs faster, it may mitigate some of the safety concerns associated with integrating algorithms into the nuclear domain, because accidents with those algorithms will become less likely and the confidence of operators may increase faster. And if technological development is too slow, it similarly will make accidents less likely, because militaries will be more likely to think the systems are not ready for deployment. The biggest risk will occur if there is enough enthusiasm about specific algorithms to prompt deployment, but the algorithms still face large-scale safety and reliability concerns that the research community has not addressed.

Similarly, there are some applications of AI that, if implemented to the maximum extent envisioned by strategists, would probably undermine aspects of second-strike stability for some countries. In this case, the slower the advances in AI, the lower the likely impact on strategic stability. For example, the argument against why AI-based surveillance will enable more counterforce strikes involves assumptions about the trajectory of advances in algorithms and other technologies such as batteries, communications, and the speed and endurance of strike assets. If AI capabilities and related technologies advance faster than most experts currently assume, those would become greater risks.

Confidence in Second-Strike Capabilities

A second factor that will influence the extent to which advances in AI impact strategic stability will be the ex-ante confidence (that is, confidence in predictions before an event occurs) that nuclear-armed states have in their second-strike capabilities, along with the importance they place on their nuclear capabilities vis-à-vis their conventional capabilities. After all, the more confident states are in their second-strike capabilities, independent of AI, the less willing they might be to consider applications of AI that have potential upsides in speed and endurance, but downsides in terms of reliability.

Confidence in existing second-strike capabilities may be one reason why senior US officials have sounded far from enthusiastic about the integration of AI into core areas of the nuclear domain. In contrast, Russian interest in uncrewed submarines and bombers could logically follow, given their perceived conventional inferiority relative to the United States, meaning they may have a relatively higher degree of reliance on their nuclear weapons earlier in a conflict than the US. This argument also suggests that, all else being equal, countries that perceive themselves as highly vulnerable nuclear powers, such as North Korea or even Pakistan, might be relatively more tempted by the potential of automation.[67] In contrast, countries that perceive themselves as having more reliable second-strike capabilities may be more likely to continue more analog nuclear processes, especially when it comes to early warning and response. This is an empirical question that we will learn more about in the years ahead.

63 Michael Horowitz, *The Diffusion of Military Power: Causes and Consequences for International Politics* (Princeton University Press, 2010); Michael C. Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power," *Texas National Security Review* 1, no. 3 (May 2018): 36–57, https://doi.org/10.15781/T2639KP49.

64 Barry Posen, *The Sources of Military Doctrine: France, Britain, and Germany Between the World Wars* (Cornell University Press, 1984); Stephen Biddle, *Military Power: Explaining Victory and Defeat in Modern Battle* (Princeton University Press, 2006); Stephen Peter Rosen, *Winning the Next War: Innovation and the Modern Military* (Cornell University Press, 1994).

65 Grace et al., "When Will AI Exceed Human Performance?"

66 Paul Scharre, "Killer Apps," *Foreign Affairs*, November 23, 2021, https://www.foreignaffairs.com/articles/2019-04-16/killer-apps.

67 Horowitz, Scharre, and Velez-Green, "A Stable Nuclear Future?" This scenario could also have implications for arms control if larger arsenals or more sophisticated launch vehicles decrease the need for more dangerous applications of AI.

Regime Type

A third factor that could affect adoption of AI applications in the nuclear domain, and thus the consequences for strategic stability, is regime type. Specifically, autocratic regimes, by definition, distrust their populations and often their military forces. Democratic nuclear powers such as the United States, United Kingdom, and Israel, even though all have high-technology militaries, regularly talk about the importance of their people, and stress that their trained and professional militaries are the key to their success.

The increasing capabilities of robotic systems and algorithms thus present an opportunity for autocratic regimes to decrease their reliance on potentially unreliable soldiers.

Autocratic regimes often face constraints on their ability to use their militaries due to the demands of coup-proofing—the need for autocratic leaders, at times, to make military decisions about promotion and strategy not based on merit or the ability to counter external threats, but based on the desire to prevent the development of internal power centers that could threaten the regime.[68] While not universally true, autocratic regimes generally have relatively smaller selectorates than do democracies,[69] and desire strategies that allow them to reduce their reliance on soldiers, who might be unreliable unless chosen from the most loyal population centers.

The increasing capabilities of robotic systems and algorithms thus present an opportunity for autocratic regimes to decrease their reliance on potentially unreliable soldiers. Today, deploying a fighter squadron requires putting pilots in the cockpit and deploying them on the battlefield with weapons that could, in theory, be turned against the regime. In a world of remotely piloted systems or systems enabled by AI, this scenario becomes less of a concern. Loyalists could operate larger, robotic units from centralized command posts that the regime can control. Thus, while initially autocratic regimes might appear skeptical of AI due to their desire to centralize the use of military force—meaning they would not want to delegate power to machines—all else being equal, their distrust of their own people could lead them to an ever-larger embrace of AI.[70] The logic of AI-enabled capabilities would then cut against the way that autocracies attempting to coup-proof their militaries often have multiple military or military-like organizations to counterbalance each other.[71]

What does this mean in the nuclear domain? Autocratic regimes armed with nuclear weapons are likely to want to decrease the risk that their soldiers would disobey an order to use nuclear weapons because of hidden opposition to the regime. Autocratic leaders might also view algorithms as a way to centralize decision-making over nuclear use. The fewer people in between the autocrat and nuclear launch, the more centralized the authority, after all. Thus, autocratic regimes, for reasons that have to do with their own fear of losing office and distrust of their own populations, might be more risk-acceptant when it comes to dangerous uses of AI within their nuclear infrastructure.

Assumptions About the Probability of Accidents

A final factor that might influence whether uses of AI undermine strategic stability has to do with assumptions about the validity of normal accident theory.[72] To what extent will the integration of AI into different aspects of the nuclear domain increase the complexity and coupling of systems in a way that makes accidents less likely?

This question is particularly relevant given the variety of failure modes possible with algorithms. Inadequate training data or coding can create reliability challenges that make failures more likely. The inability of algorithms to operate effectively most of

68 Ulrich Pilster and Tobias Böhmelt, "Coup-Proofing and Military Effectiveness in Interstate Wars, 1967–99," *Conflict Management and Peace Science* 28, no. 4 (September 1, 2011): 331–50, https://doi.org/10.1177/0738894211413062; Caitlin Talmadge, *The Dictator's Army: Battlefield Effectiveness in Authoritarian Regimes* (Cornell University Press, 2015); Cameron S. Brown, Christopher J. Fariss, and R. Blake McMahon, "Recouping After Coup-Proofing: Compromised Military Effectiveness and Strategic Substitution," *International Interactions* 42, no. 1 (2016): 1–30.

69 Bruce Bueno de Mesquita et al., "Testing Novel Implications from the Selectorate Theory of War," *World Politics* 56, no. 3 (2004): 363–88.

70 Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power."

71 On the competitive logic in autocracies, see Sheena Chestnut Greitens, *Dictators and Their Secret Police: Coercive Institutions and State Violence* (Cambridge University Press, 2016). Leninist regimes such as China tend not to follow those counterbalancing tendencies, meaning they would be even more prone to the centralization logic described above.

72 Sagan, *The Limits of Safety.*

the time outside the context of their programming means that, if presented with, for example, an early-warning situation where there was not sufficient training data for the algorithm, unpredictable errors become more likely.[73] Moreover, adversaries may have incentives to poison training data or hack into systems to distort algorithms, which can also make accidents more likely and further decrease reliability.[74]

Part of the question about AI and the probability of accidents, then, may be a question of timing. In a trust gap period, countries are unlikely to deploy AI systems without strong checks on reliability with regard to accidents. But in an overconfidence period, as automation bias creeps in, the dangers of automation will grow for other reasons.

Normal accident theory raises the question of whether any attempts to shift nuclear systems toward a more digital infrastructure (thereby increasing their complexity) inherently increase the risk of accidents—and by how much.[75] If one believes that the avoidance of accidental nuclear war during the Cold War came mostly from luck, rather than skill and careful planning, integration of AI can be inherently risky,[76] excluding a situation where an algorithm is as reliable and explainable as a calculator (which is unlikely, since that would represent an automated system, not an algorithm).[77] Alternatively, if one believes that it is possible, through planning and especially the education and training of human operators, to effectively operate AI systems, accident risks may seem more manageable. Regardless, additional education and training regarding the strengths and weaknesses of algorithms, and ensuring that operators have specific knowledge about the limits of any algorithms employed in the nuclear domain, seem like plausible and constructive policy steps.[78]

Confidence-Building Measures and the Potential for Risk Reduction

Given that there are clear risks for strategic stability associated with the integration of AI into the nuclear domain even in the best case, what are strategies to potentially manage or mitigate those risks? One possibility is a tool developed during the Cold War to make inadvertent war less likely: confidence-building measures, which are actions to reduce the risk of inadvertent war that can arise from military competition. CBMs include measures such as transparency, notification, and monitoring. As Marie-France Desjardins describes,[79] the four most common facets of confidence-building measures are:

- information sharing and communication,
- measures to allow for inspections and observers,
- "rules of the road" to govern military operations, and
- limits on military readiness and operations.

CBMs are generally thought of as less formal than arms control, since they serve as building blocks for trust building and potential cooperation. They rely on the idea that a shared interest in avoiding inadvertent war means that even competitors have potential areas of cooperation, just as the United States and Soviet Union did during the Cold War.[80]

CBMs seem potentially attractive in the military AI context because even competitors such as the United States and China do have shared interests in ensuring that, if a conflict escalates, an accident will not trigger an escalatory spiral. (There are important questions that need to be answered more broadly concerning how countries signal with AI capabilities.[81]) Ongoing Track II dialogues on AI safety

73 Dario Amodei et al., "Concrete Problems in AI Safety," *arXiv*, July 25, 2016, http://arxiv.org/abs/1606.06565.

74 Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Future of Humanity Institute, Oxford University, February 2018; Anh Nguyen, Jason Yosinski, and Jeff Clune, "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images," *arXiv*, 2014, https://doi.org/10.48550/ARXIV.1412.1897; Alejandro Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," *Information Fusion* 58 (2020): 82–115; Miles Brundage et al., "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims," *arXiv*, April 20, 2020, http://arxiv.org/abs/2004.07213.

75 Perrow, *Normal Accidents*.

76 Blair, "The Logic of Accidental Nuclear War"; Sagan, *The Limits of Safety*.

77 Donghee Shin, "The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI," *International Journal of Human-Computer Studies* 146 (February 1, 2021): 102551, https://doi.org/10.1016/j.ijhcs.2020.102551.

78 Horowitz and Kahn, "The AI Literacy Gap Hobbling American Officialdom."

79 Marie-France Desjardins, *Rethinking Confidence-Building Measures*, 1st ed. (Routledge, 2014), https://doi.org/10.4324/9781315000459.

80 Ralph A. Cossa, "Security Implications of Conflict in the South China Sea: Exploring Potential Triggers of Conflict" (Center for Strategic and International Security, 1998).

81 Michael C. Horowitz and Lauren Kahn, "Leading in Artificial Intelligence Through Confidence Building Measures," *The Washington Quarterly* 44, no. 4 (2021): 91–106.

and international cooperation are designed to improve mutual understanding between countries such as the United States and China; these activities serve as ends in themselves, and potentially create frameworks for further cooperation.

More concrete risk-reduction CBMs may also be worth considering, particularly if they can decrease the risk of accidents. CBMs could directly reduce the risk of inadvertent conflict arising from military applications of AI, raising the question of which CBMs might have the most utility.[82] The current willingness of the United States to consider limits at the intersection of AI and nuclear weapons—given the way the United States generally opposes capability limits—suggests one potential area for proposed agreement or unilateral declarations. Countries could agree to support positive human control for nuclear launch decisions, opposing taking a human out of the loop. Thus, even if a country integrates AI for pattern-recognition purposes into its early-warning systems or surveillance of the nuclear capabilities of other states, they could agree to keep a human in the loop when it comes to nuclear launch. The result would regulate autonomous "dead hand" systems or automatic "triggers" to use nuclear weapons.[83]

A positive nuclear-control CBM, even if just best practices or rules of the road, could make it less likely that an algorithmic error—whether due to flaws in algorithm development, scenarios the algorithm could not anticipate, or something else—leads to nuclear war. One challenge would be that countries less secure in their second-strike capabilities may be less willing to negotiate or sign such an agreement, because the advantages of automatic triggers for nuclear use may outweigh the risks, particularly if they are worried about decapitation at machine speed.

Another possibility is negotiating toward an autonomous incidents agreement, designed to mirror some of what the Incidents at Sea Agreement between the United States and the Soviet Union aimed to accomplish during the Cold War.[84] The Incidents at Sea Agreement was designed to reduce the risk that naval interactions would lead to escalation because of misunderstandings about the types of ships transiting contested spaces or the purposes of naval movements. Through notification procedures and information sharing, the Incidents at Sea Agreement decreased, albeit mildly, the risk of accidental conflict.

Countries could agree to support positive human control for nuclear launch decisions, opposing taking a human out of the loop.

The same may be possible in thinking about AI-enabled autonomous systems (since AI will be part of nearly all systems in one way or another). An agreement to encourage information sharing about these systems, especially if deployed in the field, could reduce the risk that potential accidents lead to miscalculation. Particularly given the difficulty of signaling that something is, in fact, an accident when dealing with software, an agreement may help build confidence and trust. This type of agreement would function best in peacetime. There would always be some degree of uncertainty about the level of autonomy in a system, since it is not directly observable. But if there is shared agreement on the risks of accidents with AI-enabled autonomous systems, countries could be incentivized, at least in peacetime, to notify adversaries in ways that decrease the risk of a spiraling accident.

For all of these measures, a key question is the extent to which agreements, even if informal, would require degrees of information disclosure or transparency that might reveal information about capabilities or help potential adversaries make their own AI systems more reliable and therefore more effective. Given concern about China's potential to overtake the United States in the AI arena, countries such as the US may be especially cautious when it comes to CBMs surrounding AI. The easiest CBMs may be those that involve the equivalent of permissive-action links for the twenty-first century—technology or best-practice sharing that improves safety without

82 Michael C. Horowitz, Lauren Kahn, and Casey Mahoney, "The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?," *Orbis* 64, no. 4 (January 1, 2020): 528–43, https://doi.org/10.1016/j.orbis.2020.08.003; Andrew Imbrie and Elsa B. Kania, *AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement* (Center for Security and Emerging Technology, Georgetown University, 2019).

83 The Global Commission on Responsible AI in the Military Domain has proposed legally binding limits on AI integration in nuclear command and control. See The Global Commission on Responsible Artificial Intelligence in the Military Domain, *Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of Artificial Intelligence in the Military Domain* (Hague Centre for Strategic Studies, 2025).

84 Kurt M. Campbell, "The US–Soviet Agreement on the Prevention of Dangerous Military Activities," *Security Studies* 1, no. 1 (1991): 109–31; Sean M. Lynn-Jones, "A Quiet Success for Arms Control: Preventing Incidents at Sea," *International Security* 9, no. 4 (1985): 154–84, https://doi.org/10.2307/2538545; David F. Winkler, "The Evolution and Significance of the 1972 Incidents at Sea Agreement," *Journal of Strategic Studies* 28, no. 2 (April 1, 2005): 361–77, https://doi.org/10.1080/01402390500088395.

improving capabilities. But how that could work remains unclear from a technical perspective.[85]

Conclusion

The world is very early in the age of artificial intelligence, particularly when it comes to scaled military applications, as opposed to experiments. The possibility of using AI-driven tools such as pattern-recognition algorithms for tasks such as early warning or target identification could be attractive to states in the conventional and nuclear domains. However, the multitude of possible accidents involving AI, whether from training-data limitations, data poisoning, hacking, or spoofing, will generate risks.

How significant these risks are, and how countries resolve them, will not just depend on the actual state of the technology, but on the beliefs that countries have about the technology. While underconfidence could generate trust gaps that mean countries forgo useful capabilities, overconfidence can raise the prospect of automation bias, with potentially dangerous consequences in the nuclear domain. Moreover, the prospect that conventional uses of AI could place pressure on strategic stability could loom even if countries maximize safety in AI applications in the nuclear domain. Other factors, such as technological progress and beliefs about accidents, could further shape trust and confidence in AI, impacting strategic stability more broadly.

Confidence-building measures surrounding military AI represent one potential way to take advantage of countries' shared interests in avoiding accidental war. While broader cooperation may be challenging, a CBM framework offers a potentially productive way forward.

Michael C. Horowitz *is director of Perry World House and the Richard Perry Professor at the University of Pennsylvania. He is also senior fellow for technology and innovation at the Council on Foreign Relations (CFR). From 2022 to 2024, he served as deputy assistant secretary of Defense for Force Development and Emerging Capabilities and director of the Emerging Capabilities Policy Office. He is the author of* The Diffusion of Military Power: Causes and Consequences for International Politics *(Princeton University Press, 2010), and the coauthor of* Why Leaders Fight *(Cambridge University Press, 2015). He won the Karl Deutsch Award given by the International Studies Association for early career contributions to the fields of international relations and peace research. Professor Horowitz is a life member at CFR. Professor Horowitz received his PhD in government from Harvard and his BA in political science from Emory.*

University of Pennsylvania, Philadelphia, PA, USA, email: horom@sas.upenn.edu.

Acknowledgments: *The author would like to thank the reviewers and editors for their helpful suggestions. This work was supported in part by Carnegie Corporation of New York. All errors are the responsibility of the author.*

Image: *11th Marine Expeditionary Unit by Sgt. Trent A. Henry.*[86]

85 Brundage et al., "Toward Trustworthy AI Development."

86 For image, see https://www.dvidshub.net/image/9272262/us-marines-conduct-dead-center-suas-training.