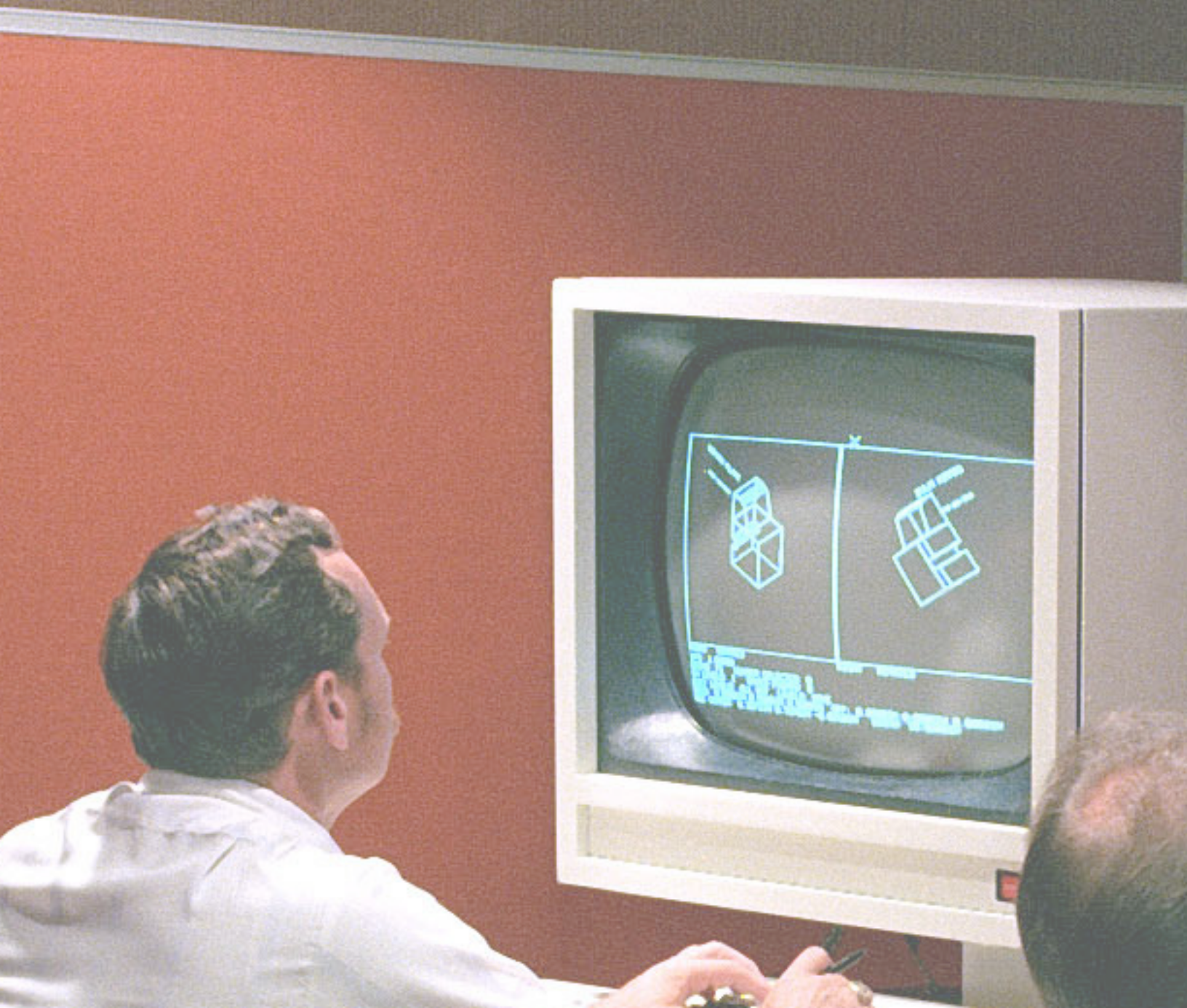
FAS FEDERATION OF AMERICAN SCIENTISTS

future of life INSTITUTE

Artificial Intelligence and Global Risk

A Framework for Policymakers



Artificial intelligence (AI) technologies are increasingly shaping domains central to U.S. policy, including economic competitiveness, national security, public administration, and scientific research.[1] AI systems offer substantial potential benefits, including improved data analysis (and fusion of disparate data flows), forecasting, and operational efficiency in government and military contexts.[2]

At the same time, increasingly capable AI systems exacerbate existing risks, including the potential for scalable offensive cyber operations, the manipulation of information environments via synthetic media, the potential proliferation of chemical, biological, nuclear, and radiological weapons, and creating new forms of escalation risks during a crisis among militaries deploying a wide variety of AI tools both to support and engage in war.

AI systems themselves may also present new risks—with some analysts suggesting the advent of AI models capable of self-guided improvement may yield systems vastly more capable than humans. Such AIs may not be controllable, and could pose risks at a global scale.[3]

THIS RISK IS DIFFERENT

Unlike earlier harbingers of global risks—such as nuclear weapons—AI is a general-purpose, "enabling" technology embedded within complex "sociotechnical" systems. It also involves a much broader set of actors—particularly those in the private sector—that complicate the legislative and regulatory response to the technology compared to historical examples that initially started within the government (e.g., nuclear weapons and the internet).

Risks associated with AI are systemic: they emerge from interactions among capabilities, institutions, and infrastructure rather than from discrete technologies or actors that previously defined the governance regimes that we sought to govern legacy technologies (e.g., nonproliferation, arms control, export control rules).

WHAT KIND OF TECHNOLOGY IS AI?

Understanding these risks requires clarity about the nature of the technology itself. Indeed, policymaking in this space is complicated by persistent disagreement about AI's trajectory, capabilities, and implications. These perspectives shape (often implicitly) how policymakers prioritize risks and proposed interventions.

01. IS IT JUST HYPE?

Some perspectives emphasize that AI is subject to hype, arguing that current capabilities are overrated, economic constraints are binding, and risks may be overstated relative to historical technological change. Those within this camp see current systems as fundamentally limited, and expect those limitations to persist.[9]

02. HOW ABOUT "SUPERINTELLIGENCE"?

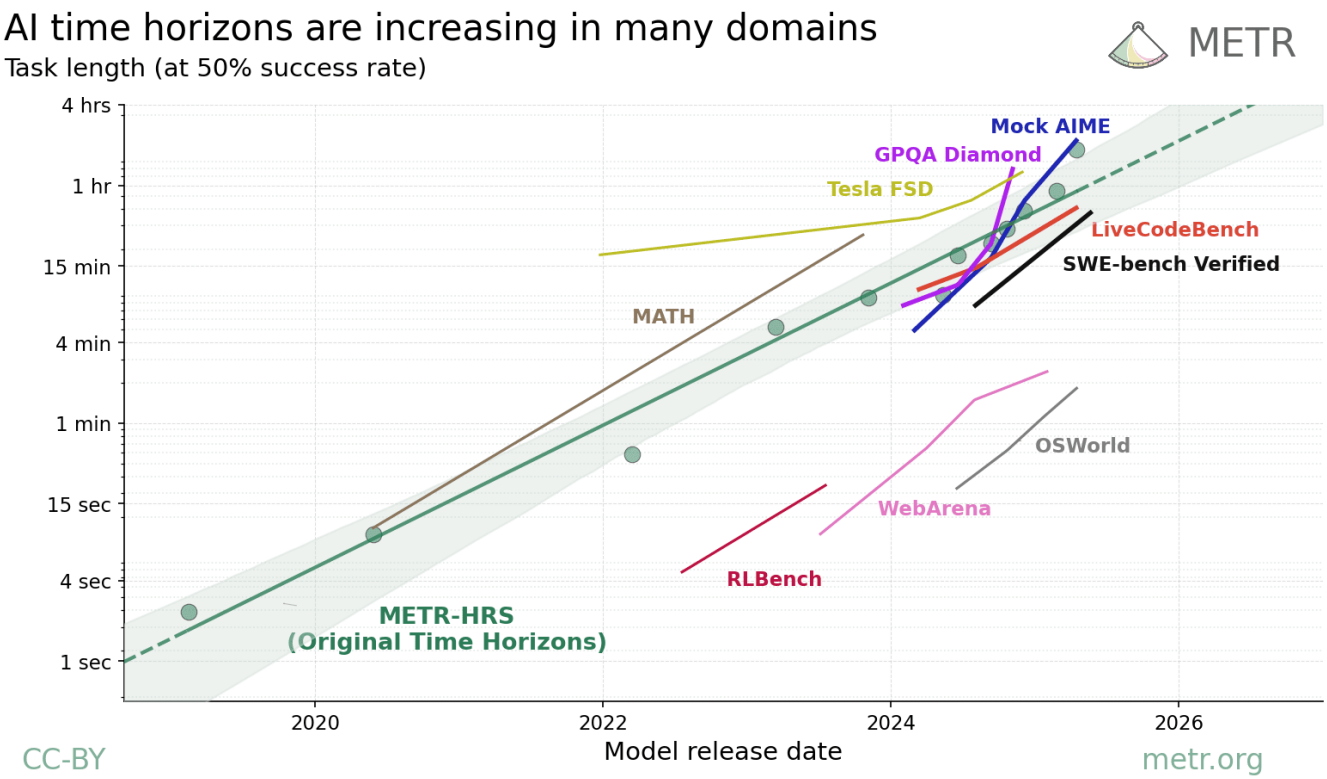
Another perspective focuses on the possibility that AI systems could become increasingly autonomous or transformative, introducing discontinuities in capability and raising concerns about long-term control over systems more capable than humans.[10]

03. ... OR IS IT NORMAL?

For others, AI is neither pure hype nor on the path to "superintelligence". It is a powerful but ultimately "normal" technology: one where the key actors are still humans and institutions, but not alone. It is the human–AI team, the firm, the bureaucracy, the military organization, the market, or the state that redesigns itself as it adopts AI over a period of decades. To view AI as normal is "not to understate its impact," but to reject the tendency to treat it as "a separate species, a highly autonomous, potentially superintelligent entity."[11]

AI Systems Today

At present, AI systems—particularly those described as "frontier models"—demonstrate a wide range of capabilities, including the ability to process large datasets, generate text and code, and support decision-making in technical and operational contexts. They are also improving rapidly.



AI CAPABILITY LIMITATIONS

However, these systems also exhibit important limitations. Their performance is often uneven, with strong results in some tasks and significant failures in others. This "jagged" capability profile complicates efforts to generalize about system performance.[4] In addition, there is often a gap between demonstrated capability in testing environments and reliable performance in real-world conditions. Further, AI capability is not linear: these capabilities depend on data inputs, environmental conditions, and the specific ways in which systems are deployed, and can feed directly into whether capabilities perform successfully or not.

AI IS CONTEXT-DEPENDENT

AI systems are also rarely deployed in isolation. Instead, they are embedded within broader institutional architectures that include human operators, processes, data pipelines, and physical and digital infrastructure. In practice, the relevant unit of analysis is often not the AI model itself, but the human–machine ecosystem. Outcomes depend not only on model performance, but on how outputs are interpreted, integrated into decision-making processes, and acted upon within organizational contexts. In our view, this broader framing is essential for understanding how risks arise and propagate.

A Framework for Global Risk

THREAT, VULNERABILITY, AND CONSEQUENCE

One framework for analyzing AI-related risks draws on a common heuristic in risk analysis often used in national security contexts, especially where risk can be understood as a function of threat, vulnerability, and consequence.

THREAT refers to who acts and why, examining the intentions and capabilities of relevant actors—from governments to terrorist groups.[12]

VULNERABILITY refers to where systems are susceptible to failure, manipulation, or misuse, including technical issues with AI systems, weaknesses in how they are adopted and governed, and larger institutional and societal fault lines.[13]

CONSEQUENCE refers to what happens when failures occur, including the scale, duration, and systemic impact of resulting harms, with AI tools often offering an accelerant to proliferation, deployment, and/or use of AI tools to induce harm across a wide variety of domains.[14]

EXAMPLE: HOW AI ACCELERATES CYBER ATTACKS

For example, in the cyber domain, AI-enabled tools can accelerate key steps in the cyber kill chain[5]—from reconnaissance to exploit development—reducing the time, cost, and expertise required for sophisticated operations. This shifts the threat landscape by pairing existing malicious intent with enhanced capability, enabling both state and non-state actors to scale attacks and operate with greater speed and coordination, even if fully autonomous operations remain limited.

Vulnerabilities stem from complex, digitally dependent systems and human susceptibility to targeted social engineering, which AI can exploit at scale. The consequences extend beyond discrete breaches: even limited incidents can cascade across interconnected systems, disrupting critical infrastructure, imposing economic costs, and increasing risks of misperception and escalation in high-stakes contexts.

How policymakers deal with these risks is shaped by broader (and often implicit) perspectives as to the nature of AI as a technology. While those that view AI technologies as subject to a hype cycle would see the concerns above as overrated, those that view AI as a normal technology might view this reality as manageable through diffusion of defensive systems over the next couple of decades, and, finally, those in the superintelligence camp might view the serious risks of autonomous AI systems executing highly sophisticated cyber attacks beyond human control in the near future requiring a revolutionary regulatory approach.

AI technologies alter the nature of global risks in several important ways. Historically, high-consequence risks associated with specific technologies controlled by a relatively small number of actors (e.g., nuclear weapons).[6] AI technologies, by contrast, operate across multiple domains simultaneously and diffuses rapidly across states, firms, and non-state actors. This diffusion lowers barriers to entry and complicates efforts to control access to capabilities.[7] At the same time, AI amplifies the scale and speed of action, enabling both beneficial and harmful activities to be carried out more efficiently. Finally, as AI systems become more complex and more deeply integrated into decision-making processes, there is the potential for a reduction in effective human control, particularly in environments characterized by time pressure and uncertainty.[8]

Key Policy Challenges

A central challenge is the problem of uncertainty and the limited empirical evidence available to guide decision-making. AI capabilities are evolving rapidly, often outpacing the accumulation of real-world data on system performance and impact, and are subject to uncertainty given the opacity of these systems. Policymakers face a tradeoff between acting early with incomplete information and waiting for more evidence while risks continue to evolve with no clear baseline as to acceptable levels of performance in high-risk settings. A related challenge concerns measurement and evaluation. Existing testing and benchmarking approaches often fail to capture how systems will perform in adversarial, dynamic, or high-stakes environments. This limits the ability to assess risk and design effective safeguards. Finally, there is a mismatch between existing, inflexible governance institutions and the cross-domain, rapidly evolving nature of AI technologies. Many institutions remain organized around specific sectors or policy areas, while AI-related risks cut across these boundaries.

In light of these challenges, several policy considerations emerge. Improving measurement and evaluation represents a clear priority, including the development of context-specific and continuous testing methods that better reflect real-world conditions and use cases. Increasing transparency—both in terms of system capabilities and deployment contexts—can improve oversight and accountability. Strengthening government capacity is also important, particularly by investing in technical expertise within public institutions and enhancing the ability to assess AI systems in the context of high-impact use cases. Finally, policymakers may seek to design policies that are robust to uncertainty, explicitly stating underlying assumptions and identifying governance approaches that perform well across a range of possible futures.

To support decision-making, policymakers may benefit from asking a set of structured questions of those proposing particular policy options across the risk space:

- ↳ Who or what are you treating as the relevant actor? (e.g., humans, human–AI systems, or autonomous AI systems themselves)
- ↳ What evidence are you relying on? (e.g., benchmarks, deployment history, historical analogies, adoption trends, measured impacts, expert forecasts)
- ↳ What kind of risk are you primarily worried about? (e.g., accident, adversarial misuse, or losing control over autonomous systems)
- ↳ What evidence would actually change your mind about a particular risk?
- ↳ Do you expect that evidence to arrive before the policy window to address that risk closes?

Looking ahead, the analysis of AI-related risks depends in part on how the technology is conceptualized—whether as pure hype, a general-purpose technology, or a potentially superintelligent system. These different perspectives shape both the assessment of risk and the design of governance strategies.

Policy Recommendations

AI-related global risks will not be managed through a single intervention. These risks emerge from the interaction of capabilities, institutions, infrastructure, and human decision-making. The risks can also change substantially depending on which trajectory AI development follows.

That means policy needs to operate in layers and be responsive to multiple possible futures. Some of the recommendations below reduce the level of threat. Others reduce vulnerability or consequence. Still others reduce the uncertainty that cuts across all three.

FOUNDATION: BUILD GOVERNMENT CAPACITY TO UNDERSTAND AND ACT

None of what follows works without baseline capacity. Governments need the ability to evaluate AI systems, interpret technical claims, coordinate across agencies, and engage credibly with private-sector developers. That includes sustained investment in technical evaluation, stronger talent pipelines, mechanisms for surge capacity in crises, and clearer norms for how AI is used inside government. Without this, policy becomes reactive and dependent on external actors.

01. STRENGTHEN TESTING, EVALUATION, VERIFICATION, AND VALIDATION

Policymakers need a clearer picture of what AI systems can do, where they fail, and how that translates into real-world risk. That means moving beyond static benchmarks toward evaluation that reflects operational conditions: reliability over time, performance under stress, and behavior in adversarial settings. Evaluation should increasingly focus on systems as they are actually used, not just models in isolation.

02. BUILD EARLY-WARNING INDICATORS FOR CAPABILITY SHIFTS

AI is evolving faster than the evidence base used to govern it. Government should track a small set of indicators that help distinguish between different trajectories: diffusion rates, autonomy over longer tasks, AI-enabled research and development, and the emergence of dual-use capabilities. These indicators should be tied to decision points in order to shorten the gap between capability change and policy response, and should also require multi-stakeholder input to provide the best-available information for decisionmakers.

03. APPLY RISK-TIERED GOVERNANCE TO HIGH-CAPABILITY SYSTEMS

Some AI systems will matter more than others. Policymakers should focus attention on models and systems with the potential to create outsized risk, particularly those with advanced capabilities, broad access, or weak safeguards. This includes measures related to model security, access controls, incident reporting, and transparency around capabilities and limitations, which would indicate that CAISI should lead or own many of these activities. The objective is to shape how powerful systems are developed and distributed before risks become harder to manage downstream.

04. GOVERN DEPLOYMENT AS A SOCIOTECHNICAL PROBLEM

Most failures will not come from models in isolation. They will come from how those models are embedded in real-world systems. High-stakes deployments should be treated accordingly. This includes requirements for independent evaluation, attention to human factors, clear accountability, and mechanisms for intervention when systems behave unexpectedly. A human in the loop is only meaningful if that human can actually exercise judgment.

05. TILT THE BALANCE TOWARD DEFENSE

AI lowers the cost of certain offensive capabilities, from cyber operations to information manipulation. It can also strengthen defense. Policy should deliberately accelerate defensive applications, including cybersecurity, biosecurity, and infrastructure protection, while ensuring that high-risk capabilities are not broadly accessible without safeguards. The goal is to give defenders time and capability to adapt.

06. INVEST IN RESILIENCE AND CONSEQUENCE MANAGEMENT

Even well-governed systems will fail. The question is how much those failures matter. Policymakers should invest in resilience: incident reporting, crisis response, redundancy, and cross-domain coordination. This is particularly important in contexts where AI can compress timelines, increase complexity, or amplify the effects of error.

Taken together, these recommendations are designed to hold across different views of AI's future. They improve decision-making under uncertainty, strengthen institutions, and reduce the likelihood that capability gains translate into large-scale harm.

This briefing booklet is authored by Andrew Reddie, PhD, Oliver Stephenson, PhD, and Yong-Bee Lim, PhD. A wider FAS-FLI report, 'Converging Risks: AI and the Future of Global Security', will be published on May 15th, 2026 on fas.org.

Endnotes

1 "AIx Compendium Report: Converging Risks: AI and the Future of Global Security," Federation of American Scientists, 2026. https://www.crs.gov/Reports/R47644; https://www.crs.gov/Reports/R46795

2 Schmidt, Eric, Robert Work, Safra Catz, Eric Horovitz, Steve Chien, Andrew Jassy, Mignon Clyburn et al. "National security commission on artificial intelligence (ai)." (2021).

3 https://arxiv.org/pdf/2306.12001

4 Gans, Joshua S. A Model of Artificial Jagged Intelligence. No. w34712. National Bureau of Economic Research, 2026; https://helentoner.substack.com/p/taking-jaggedness-seriously.

5 https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html.

6 https://fas.org/wp-content/uploads/2026/01/January-2026-AI-Bio.pdf

7 https://fas.org/wp-content/uploads/2025/12/1215-ai-mil.pdf

8 https://fas.org/wp-content/uploads/2025/07/June2025_AIxNC3_FAS.pdf

9 Hanna, Alex. "The AI Con: How to Fight Big Tech's Hype and Create the Future We Want." (2025).

10 https://www.cfr.org/articles/artificial-intelligence-is-facing-a-crisis-of-control-and-the-industry-knows-it; https://superintelligence-statement.org/; https://ai-2027.com/.

11 https://knightcolumbia.org/content/ai-as-normal-technology

12 https://unicri.org/News/Algorithms-Terrorism-Malicious-Use-Artificial-Intelligence-Terrorist-Purposes; Johnson, James. "The AI-cyber nexus: implications for military escalation, deterrence and strategic stability." Journal of Cyber Policy 4, no. 3 (2019): 442-460; Bontridder, Noémi, and Yves Poullet. "The role of artificial intelligence in disinformation." Data & Policy 3 (2021); Helmus, Todd C. "Artificial intelligence, deepfakes, and disinformation: A primer." (2022); Mayer, Michael. "Trusting machine intelligence: artificial intelligence and human-autonomy teaming in military operations." Defense & Security Analysis 39, no. 4 (2023): 521-538.

13 Erskine, Toni, and Jenny L. Davis. ""Borgs in the org" and the decision to wage war: The impact of AI on institutional learning and the exercise of restraint." In Cambridge Forum on AI: Law and Governance, vol. 1, p. e45. Cambridge University Press, 2025; Jensen, Benjamin M., Christopher Whyte, and Scott Cuomo. "Algorithms at war: the promise, peril, and limits of artificial intelligence." International Studies Review 22, no. 3 (2020): 526-550; Horowitz, Michael C., 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. 2 (2024): sqae020; See also, https://futureoflife.org/project/artificial-escalation/; Raska, Michael, and Richard A. Bidiffusetzinger, eds. The AI wave in defence innovation: Assessing military artificial intelligence strategies, capabilities, and trajectories. Taylor & Francis, 2023.

14 Horowitz, Michael C. "Artificial Intelligence and the Future of Strategic Stability." Texas National Security Review. Vol 9. no. 2 (2026). https://doi.org/10.1353/tns.00034.