FAS FEDERATION OF AMERICAN SCIENTISTS

MAY 2026

Converging Risks

AI and the Future of Global Security

About FAS

The **Federation of American Scientists (FAS)** is an independent, nonpartisan think tank that brings together members of the science and policy communities to collaborate on mitigating global catastrophic threats. Founded in November 1945 as the Federation of Atomic Scientists by scientists who built the first atomic bombs during the Manhattan Project, FAS is devoted to the belief that scientists, engineers, and other technically trained people have the ethical obligation to ensure that the technological fruits of their intellect and labor are applied to the benefit of humankind. In 1946, FAS rebranded as the Federation of American Scientists to broaden its focus to prevent global catastrophes.

Since its founding, FAS has served as an influential source of information and rigorous, evidence-based analysis of issues related to national security. Specifically, FAS works to reduce the spread and number of nuclear weapons, prevent nuclear and radiological terrorism, promote high standards for the safety and security of nuclear energy, illuminate government secrecy practices, and prevent the use of biological and chemical weapons.

FAS can be reached at 1150 18th St. NW, Suite 1000, Washington, DC, 20036, **fas@fas.org**, or through **fas.org.**

COVER: CONTROL DATA PROCESSING IBM EQUIPMENT OAK RIDGE TENNESSEE MARCH 1967
VIA DEPARTMENT OF ENERGY

The Global Risk Program at FAS

The Global Risk Program at the Federation of American Scientists (FAS) focuses on addressing and preventing the events and threats that could permanently cripple or destroy humanity. Some key areas our team focuses on include nuclear war, the next global pandemic, biological attack, and even a collision with a massive near-earth object. Our team of policy experts, scientists, and researchers use tools including forecasting, research, and analysis, and expertise in key global risk domain areas to develop modern policy solutions for a rapidly advancing and complex time in humanity's development. Humanity must proactively develop and pursue sound policies to protect against these dangers, including through global cooperation. Find out more at our website, www.fas.org/issue/global-risk.

FAS can be reached at 1150 18th St. NW, Suite 1000, Washington, DC, 20036, fas@fas.org, or through fas.org.

Funding

This report and the associated convenings were made possible through the generous support of the Future of Life Institute (FLI). The views expressed in this report are those of the authors and do not necessarily reflect the positions of the funders or participants.

Acknowledgements

This project is led by Dr. Yong-Bee Lim, who is the Associate Director of Global Risk at the Federation of American Scientists.

Special thanks go to the core FAS team working on this project, which includes: 1) Dr. Oliver Stephenson, Associate Director for Artificial Intelligence and Emerging Technology Policy; 2) Mr. Elliott Gunnell, Project Associate for Global Risk; 3) Ms. Angela Kellett, Senior Comms Associate; 4) Ms. Katie McCaskey, Communications Manager; 5) Ms. Kate Kohn, Senior Communications Manager; 6) Mr. Jonathan Wilson, Associate Director of Communications; and 7) Mr. Gil Ruiz, Director of Government Affairs. We also extend enormous gratitude to Dr. Andrew Reddie, the founder and director of the Berkeley Risk and Security Lab, for collaborating with us as our Senior Advisor on this report and other project activities.

Further, this project could not have been completed without everyone at FAS that supported this important work, with special attention paid to the Nuclear Information Project Team (Hans, Matt, Eliana, and MacKenzie), and the Artificial Intelligence and Emerging Technology Policy Team (Clara, Caroline, and David). We also appreciate the support of FAS Senior Leadership, including CEO Daniel Correa, CSO Jedidah Isler, PhD, and COS Annette Germana, throughout this project's tenure.

In addition, we greatly appreciate the partnership with FLI, and want to particularly thank FLI Executive Director Anthony Aguirre, PhD, Mr. Hamza Chaudhry, and FLI U.S. Communications Manager Mr. Chase Hardin for their partnership throughout this effort.

Deep appreciation also goes to Mr. Jon Wolfsthal, former Director of the Global Risk portfolio at FAS, for his leadership during his tenure on the project.

Finally, this effort would not be possible without the time and perspectives of over 300 people who attended our workshops and helped inform and shape our work. Those who consented to a public acknowledgement of their role in this work are included in a "Contributors" section (Appendix B) at the end of this report.

Contents

ABOUT FAS I

THE GLOBAL RISK PROGRAM AT FAS II

FUNDING II

ACKNOWLEDGEMENTS II

EXECUTIVE SUMMARY 1

HOW POLICYMAKERS SHOULD USE THIS REPORT - FIVE QUESTIONS, A SOLID FOUNDATION, AND FIVE PILLARS 2

CHAPTER 1. WHAT KIND OF TECHNOLOGY IS AI? 3

PERFORMANCE OF TODAY'S AI SYSTEMS 4

EVALUATING AI SYSTEMS 6

THREE VIEWS OF AI'S TRAJECTORY 7

WHY THE CAMPS DIVERGE 10

POLICYMAKERS MUST ACT UNDER UNCERTAINTY 11

CHAPTER 2. GLOBAL RISKS AND ARTIFICIAL INTELLIGENCE 13

THREAT, VULNERABILITY, AND CONSEQUENCE - A WORKING MODEL OF RISK 14

AI AS A SOURCE OF THREAT: CAPABILITY, INTENT, AND CONTROL 16

RISK, IN THEORY 17

FROM SYSTEMS TO ACTORS 19

CHAPTER 3. THREAT - INTENT AND CAPABILITY 20

ACTORS AND INTENT: WHO GENERATES AI-RELATED THREATS? 20

AI AS A THREAT ACTOR 23

THE DRIVERS OF THREAT: HETEROGENEOUS CAPABILITIES 24

REFRAMING "THREAT" FOR THE AI ERA 26

CHAPTER 4. VULNERABILITY - AN ARCHITECTURE OF FRAGILITY 28

CONTROL, VISIBILITY, AND RESTRAINT 28

TECHNICAL VULNERABILITIES 29

INSTITUTIONAL VULNERABILITIES 32
WHEN VULNERABILITY GENERATES THREAT 35
CHAPTER 5. CONSEQUENCE - A VANISHING MARGIN FOR ERROR....37
A TAXONOMY OF HARM 37
STRATEGIC (IN)STABILITY 38
WORST CASE SCENARIOS 42
DISTRIBUTION OF HARM 46
POLICYMAKER PATHOLOGIES 46
FROM THREAT TO VULNERABILITY TO CONSEQUENCE 47
CHAPTER 6. POLICY RECOMMENDATIONS....48
THE FOUNDATION. BUILD GOVERNMENT CAPACITY, COORDINATION, AND TRANSLATION INFRASTRUCTURE 48
PILLAR 1. BUILD TESTING, EVALUATION, VERIFICATION, AND VALIDATION (TEVV) CAPACITY AND EARLY-WARNING SYSTEMS FOR DANGEROUS AI CAPABILITIES 50
PILLAR 2. GOVERN THE TECHNICAL LAYER: COMPUTE, MODELS, WEIGHTS, ACCESS, SAFEGUARDS, AND DANGEROUS CAPABILITIES 52
PILLAR 3. GOVERN DEPLOYMENT IN SOCIOTECHNICAL SYSTEMS, NOT JUST MODELS 53
PILLAR 4. SHIFT THE OFFENSE-DEFENSE BALANCE OF AI SYSTEMS TOWARD DEFENSE 54
PILLAR 5. BUILD SOCIETAL AND INSTITUTIONAL RESILIENCE FOR WHEN PREVENTION FAILS 55
PUTTING THE PILLARS TOGETHER 56
SUMMARY. MAPPING THE FIVE PILLARS AND FOUNDATION ONTO THREAT, VULNERABILITY, AND CONSEQUENCE....57
FOUNDATION. GOVERNMENT CAPACITY, COORDINATION, AND TRANSLATION INFRASTRUCTURE 57
PILLAR 1. MEASUREMENT, TEVV, AND EARLY WARNING 58
PILLAR 2. TECHNICAL-LAYER GOVERNANCE 59
PILLAR 3. SOCIOTECHNICAL DEPLOYMENT GOVERNANCE 60
PILLAR 4. DEFENSIVE ADVANTAGE 61

PILLAR 5. RESILIENCE 62

CONCLUSION.....63

WHAT THIS MEANS FOR POLICYMAKERS 63

WHERE RISK REDUCTION IS MOST TRACTABLE 64

APPENDIX A. SUMMARY OF PROJECT.....66

ORIGINS AND PURPOSE 66

CONVENING AND ENGAGEMENT PROCESS 66

RECURRING THEMES AND RELATED MATERIALS 67

APPENDIX B. CONTRIBUTORS 68

Executive Summary

Artificial intelligence (AI) is no longer a standalone technology policy issue. It is becoming a general-purpose capability embedded in domains central to global security. As AI systems enter biological research, cyber operations, nuclear stability, military decision-making, and other security contexts, they are changing how global risks emerge, spread, and interact.

This report provides an evidence-based foundation for how policymakers, national security practitioners, technical experts, funders, and civil society leaders should think about the convergence of AI and global risks. It builds on a series of convenings by the Federation of American Scientists (FAS) and Future of Life Institute (FLI) focused on AI and biosecurity, cyber, nuclear risk, and military integration. Across those conversations, a common theme emerged: AI risk does not sit within any single domain or threat actor. It emerges from the interactions between increasingly capable tools and the institutions and infrastructures they operate through.

Rather than predicting a single future for AI, this report aims to help decision-makers navigate uncertainty across multiple trajectories. It recommends policies that can reduce uncertainty and remain robust across a range of possible futures.

This report focuses on general-purpose "frontier" AI systems: highly capable systems that can support many kinds of work, including analysis, coding, planning, scientific reasoning, tool use, synthetic media generation, and autonomous workflows. As of May 2026, leading frontier systems can synthesize and query large bodies of text, write and debug software, analyze technical and scientific materials, generate realistic synthetic media, and help users plan multi-step scientific or operational tasks, though their performance remains uneven and context-dependent. These capabilities are dual-use. A system that helps a researcher analyze a biological dataset may also lower barriers to harmful experimentation. A system that helps defenders identify cyber vulnerabilities may also help attackers exploit them faster.

The report starts with the three broad views of AI's future trajectory that shape policy discourse today. The "mirage" perspective sees today's AI discourse as overhyped and focuses on risks such as premature deployment, fraud, capital misallocation, and policymaker distraction. The "normal technology" view treats AI as a powerful but ultimately manageable general-purpose technology that requires serious planning, governance, and institutional adaptation. Under this view, the main risks come from uneven diffusion, brittle deployment, automation bias, and the expansion of capability to a wider set of actors. Finally, the "autonomous power" perspective argues that rapid advances in current systems may point toward increasingly autonomous or superhuman systems. From this perspective, the main risks include power concentration and loss of control over systems far more powerful than humans, with potentially existential consequences.

This report examines AI's impact on global risk largely through the "normal technology" and "autonomous power" lenses. While the "mirage" view remains important because it cautions against hype for specific AI applications, the global-risk questions at the center of this report are most visible when viewing AI as either: 1) a powerful dual-use technology diffusing through fallible institutions or; 2) a pathway toward more autonomous and powerful systems that may become harder to monitor, constrain, or control. Policymakers will need to make decisions under conditions of uncertainty. AI capabilities are evolving quickly in a period of geopolitical tension, and waiting for definitive evidence before acting may itself carry risks.

This report uses a familiar national security framework as part of the analysis: **RISK = THREAT × VULNERABILITY × CONSEQUENCE (TVC)**. Threat refers to the actors, intentions, and capabilities that generate pathways to harm. Vulnerability refers to weaknesses in technical systems, institutions, infrastructure, human-machine teams, or governance arrangements that allow threats to manifest. Consequence refers to the harms that result when threats exploit vulnerabilities, including casualties, escalation, systemic disruption, loss of trust, or long-term institutional damage.

AI may affect all three components at once. AI may increase what malicious state and non-state actors can do. It also introduces complexity to opaque systems, which increases vulnerability to something slipping through the cracks. It may also compress response timelines and make failures harder to address, which increases consequence. In autonomous power scenarios, the boundary between threat and vulnerability may blur: vulnerabilities in the oversight of powerful AI systems could result in a loss of control, resulting in AI systems that themselves pose a threat.

How Policymakers Should Use This Report - Five Questions, A Solid Foundation, and Five Pillars

As policymakers grapple with uncertain futures and rapidly advancing capabilities, we recommend these guiding questions as a way to both reduce uncertainty and surface assumptions:

- First, who or what does this proposal treat as the relevant actor: humans, human–AI systems, or AI systems themselves?
- Second, what arguments, evidence, and historical reference classes does this proposal rely on (implicitly or explicitly)?
- Third, what kind of risk or opportunity is this proposal particularly focused on?
- Fourth, what evidence would challenge or support the existence of these risks or opportunities?
- Fifth, is that evidence likely to arrive in time for policymakers to update course before the relevant risks or benefits are locked in, and can the evidence collection be accelerated?

Policymakers should also build policies that create layered defenses to reduce threat, vulnerability, and consequence. We frame policy options around a foundation and five pillars:

- **FOUNDATION.** Government capacity, coordination, and translation infrastructure. Agencies need technical expertise, access to tools, trusted channels with AI developers, cross-domain coordination, and the ability to evaluate claims about AI capabilities without relying entirely on private-sector assurances.
- **PILLAR 1.** Stronger testing, evaluation, verification, and validation (TEVV). This requires better evaluations, construct-valid benchmarks, post-deployment monitoring, and indicators for changes in AI's trajectory, including AI-enabled R&D automation.
- **PILLAR 2.** Robust technical layer governance. This includes model security, access controls, transparency obligations, incident reporting, and mitigating capabilities that create disproportionate risks.
- **PILLAR 3.** Thoughtful AI deployment in institutions and high-consequence systems. High-risk uses, such as military decision support, cyber operations, biological design workflows, and nuclear-adjacent systems, require risk-tiered approval, independent review, audit logs, human-factors testing, and other interventions to ensure meaningful human control.
- **PILLAR 4.** Shifting the offense-defense balance of novel capabilities. This includes delaying and restricting broad access to capabilities that advantage attackers while strengthening AI use in cyber defense and other defense settings.
- **PILLAR 5.** Building resilience. While resilience should not excuse preventable upstream risks, governments and institutions will need stronger cyber resilience, public health plans, and crisis communication protocols to manage future threats in these domains.

AI is already consequential, but its future trajectory remains contested. Policymakers should make their assumptions explicit, focus on what can be shaped rather than what can be perfectly predicted, and build institutions that can learn and respond as evidence changes.

Chapter 1. What Kind of Technology Is AI?

Artificial intelligence is no longer a niche issue in technology policy. Governments increasingly see it as a source of economic power, military advantage, scientific capability, and geopolitical influence. Supporters of AI point to its potential to accelerate scientific discovery, improve decision-making, and help institutions operate more effectively. However, many of the same capabilities could also scale cyberattacks, spread dangerous knowledge, increase surveillance capacity, or make complex systems harder to control.[1]

Many emerging technologies have, like AI, carried both civilian and security implications. However, AI presents a particularly difficult governance challenge because of the speed of its development, the breadth of its applications, and the uncertainty surrounding its future trajectory. Further, AI cuts across communities that think about risk in very different ways. Economic policymakers may focus on productivity and competitiveness, while national security officials may focus on strategic stability and military advantage. Cybersecurity and biosecurity experts may ask how AI changes barriers to misuse, while developers may prioritize capability gains and deployment. Civil society groups may focus on other domains, such as accountability, labor impacts, or civil rights. As a result, current AI policy debates often involve participants who operate from very different assumptions about risk, governance, and even what AI fundamentally represents.

Broadly speaking, current debates about AI tend to fall into three camps. One view holds that today's systems are overhyped, economically unstable, and unlikely to meet the expectations surrounding them. A second sees AI as a powerful but ultimately "normal" general-purpose technology: something comparable to earlier technologies such as steam or electricity. Under this view, the central challenges for technology adoption include diffusion, institutional adaptation, and uneven deployment. Finally, a third perspective argues that current systems may be early precursors to far more autonomous and capable systems, potentially including systems that escape human control. Each of these perspectives implies a very different understanding of the opportunities, risks, and governance priorities associated with AI.

These models matter because they shape what policymakers see as urgent, tractable, or even relevant. Efforts to police fraudulent AI claims and protect against financial bubbles embed one vision. Policies focused on diffusion and institutional redesign assume another. Priorities such as frontier AI control, compute governance, and data center security assume yet another. Policymakers cannot avoid making assumptions about AI's future trajectory.[2] The key governance question is whether those assumptions remain implicit within policy proposals or are made explicit enough to test, debate, and update over time.

This chapter aims to make those assumptions explicit. It maps the main views of AI shaping current policy debates, clarifies the assumptions and analogies that underlie them, and explains why these disagreements matter for assessing global risk and designing governance regimes. By making those disagreements more explicit, policymakers can better identify the assumptions guiding their decisions, the evidence that could shift their views, and the policy approaches that remain useful across multiple futures.

These views should not be reduced to simple optimism or pessimism. This report takes seriously the possibility that AI could contribute to severe or even catastrophic harms, while avoiding the assumption that such outcomes are inevitable. The core disagreement instead concerns the pace and direction of AI development: whether current

1 Jonathan B. Tucker, ed., Innovation, Dual Use, and Security: Managing the Risks of Emerging Biological and Chemical Technologies (Cambridge, MA: MIT Press, 2012).

2 Richard Danzig, Driving in the Dark: Ten Propositions About Prediction and National Security (Washington, DC: Center for a New American Security, October 26, 2011), https://www.cnas.org/publications/reports/driving-in-the-dark-ten-propositions-about-prediction-and-national-security; Andrew J. Lohn, "Beyond P(doom) for AI Risk: Quantifying Uncertainty Without Probability," Center for Security and Emerging Technology, May 2026, https://cset.georgetown.edu/publication/beyond-pdoom-for-ai-risk-quantifying-uncertainty-without-probability/.

systems are likely to disappoint, diffuse gradually through institutions, or evolve into far more autonomous and capable systems. Each perspective, therefore, carries different expectations about both opportunity and risk.

Performance of today's AI systems

"AI" can refer to a wide range of systems, from narrow classification tools to large-scale frontier models trained on enormous amounts of data and in industrial-scale facilities. This report focuses on general-purpose frontier AI systems: highly capable models that can perform many different tasks and increasingly operate with a degree of autonomy when paired with external tools and software systems.[3]

At a high level, frontier AI systems depend on three components: algorithms, data, and computing power.[4] Algorithms shape how systems identify patterns, follow instructions, reason through problems, and use tools. Data provides the text, code, images, and other examples from which models learn. And computing power, especially advanced chips in large-scale data centers, makes it possible to train and operate models at frontier scale and capability.[5]

Progress in AI typically comes from advances across all three areas: improved algorithms, larger or higher-quality datasets, and more abundant or efficient compute.[6] After training, these systems are often further adapted through fine-tuning, reinforcement learning, retrieval tools, safety filters, and agentic scaffolding that allow them to interact with software, databases, users, and external environments. See Table 1 for a list of terms relevant to this report.

This picture remains incomplete without accounting for the sociotechnical systems in which AI is deployed: the interfaces through which people use it, the workflows into which it is integrated, the incentives shaping reliance on it, and the institutions responsible for overseeing it.[7],[8] In most consequential real-world settings, the relevant actor is not the model alone, but a human using the model or an organization redesigning workflows around it. The human-AI system is, therefore, often the right unit of analysis. In addition, performance and risk depend not only on model capabilities, but also on how the system is used and governed by humans.

As of May 2026, frontier systems demonstrate growing capabilities across global risk domains.[9] They can summarize and manipulate large volumes of text, assist with coding, analyze technical literature, generate synthetic media, and, in some settings, support scientific or operational problem-solving.

3 Markus Anderljung et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety," arXiv, 2023, https://doi.org/10.48550/arXiv.2307.03718; Department for Science, Innovation and Technology, "Frontier AI: Capabilities and Risks – Discussion Paper," GOV.UK, updated April 28, 2025, https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/frontier-ai-capabilities-and-risks-discussion-paper.

4 Ben Buchanan, The AI Triad and What It Means for National Security Strategy (Center for Security and Emerging Technology, August 2020), https://doi.org/10.51593/20200021.

5 Girish Sastry et al., "Computing Power and the Governance of Artificial Intelligence," arXiv, February 13, 2024, https://doi.org/10.48550/arXiv.2402.08797.

6 Epoch AI, "Trends in Artificial Intelligence," updated February 5, 2026, accessed April 29, 2026, https://epoch.ai/trends

7 Gordon Baxter and Ian Sommerville, "Socio-Technical Systems: From Design Methods to Systems Engineering," Interacting with Computers 23, no. 1 (January 2011): 4–17, https://doi.org/10.1016/j.intcom.2010.07.003.

8 Miranda Bogen and Amy Winecoff, "Applying Sociotechnical Approaches to AI Governance in Practice," Center for Democracy and Technology, May 15, 2024, https://cdt.org/insights/applying-sociotechnical-approaches-to-ai-governance-in-practice/; Inioluwa Deborah Raji and Roel Dobbe, "Concrete Problems in AI Safety, Revisited," ICLR Workshop on Machine Learning in the Real World, 2020, arXiv preprint, submitted December 18, 2023, https://doi.org/10.48550/arXiv.2401.10899.

9 Dan Hendrycks, Mantas Mazeika, and Thomas Woodside, "An Overview of Catastrophic AI Risks," arXiv, last revised October 9, 2023, https://doi.org/10.48550/arXiv.2306.12001.

TABLE 1. COMMON TERMINOLOGY

TERM	DEFINITION
NARROW AI	AI SYSTEMS DESIGNED OR TRAINED FOR A SPECIFIC TASK, SUCH AS IMAGE CLASSIFICATION, SPAM DETECTION, OR ROUTE OPTIMIZATION.
GENERATIVE AI	AI SYSTEMS THAT PRODUCE NEW CONTENT, INCLUDING TEXT, CODE, IMAGES, AUDIO, VIDEO, OR SYNTHETIC DATA.
GENERAL-PURPOSE AI	AI SYSTEMS THAT CAN PERFORM A WIDE RANGE OF TASKS ACROSS DOMAINS RATHER THAN BEING LIMITED TO ONE NARROW FUNCTION.
FRONTIER AI	THE MOST CAPABLE AI SYSTEMS AVAILABLE AT A GIVEN TIME. THE FRONTIER CHANGES RAPIDLY AND CAN BE DIFFICULT TO DEFINE BECAUSE CAPABILITIES ARE UNEVEN, HARD TO MEASURE, AND CONTEXT-DEPENDENT.
GENERAL-PURPOSE FRONTIER AI	THE SUBSET OF FRONTIER SYSTEMS MOST RELEVANT TO THIS REPORT: HIGHLY CAPABLE MODELS AND SYSTEMS THAT CAN SUPPORT MANY KINDS OF WORK, INCLUDING ANALYSIS, CODING, PLANNING, SCIENTIFIC REASONING, TOOL USE, AND POTENTIALLY AUTONOMOUS WORKFLOWS.
FOUNDATION MODEL	A LARGE MODEL TRAINED ON BROAD DATASETS AND ADAPTED FOR MANY DOWNSTREAM USES. ITS RISKS DEPEND NOT ONLY ON THE MODEL ITSELF, BUT ON HOW IT IS DEPLOYED AND WHO CAN ACCESS IT.
LARGE LANGUAGE MODEL (LLM)	A MODEL TRAINED PRIMARILY ON LANGUAGE AND CODE THAT CAN GENERATE, SUMMARIZE, TRANSLATE, REASON OVER, OR MANIPULATE TEXT. MANY MODERN LLMS ARE ALSO MULTIMODAL OR TOOL-USING.
MULTIMODAL AI	AI SYSTEMS THAT CAN PROCESS OR GENERATE MULTIPLE KINDS OF DATA, SUCH AS TEXT, IMAGES, AUDIO, VIDEO, CODE, OR SENSOR INPUTS.
AI AGENT	AN AI SYSTEM THAT CAN TAKE ACTIONS OVER MULTIPLE STEPS, OFTEN USING TOOLS AND FEEDBACK LOOPS TO PURSUE A GOAL.
AUTONOMY	THE DEGREE TO WHICH AN AI SYSTEM CAN ACT WITHOUT DIRECT HUMAN INSTRUCTION AT EACH STEP.
MODEL WEIGHTS	THE LEARNED NUMERICAL PARAMETERS THAT ENCODE A MODEL'S CAPABILITIES. THEY MATTER BECAUSE ACCESS TO WEIGHTS CAN ALLOW OTHERS TO COPY, MODIFY, OR MISUSE POWERFUL SYSTEMS.
COMPUTE	THE SPECIALIZED COMPUTING POWER USED TO TRAIN AND RUN AI MODELS, INCLUDING ADVANCED CHIPS AND LARGE-SCALE DATA CENTERS.
AI EVALUATION / TEVV	TESTING, EVALUATION, VERIFICATION, AND VALIDATION PROCESSES USED TO ASSESS WHAT AI SYSTEMS CAN DO, HOW RELIABLY THEY PERFORM, AND WHETHER THEY ARE FIT FOR A SPECIFIC PURPOSE.
BENCHMARK	A STANDARDIZED TEST OR EVALUATION TASK USED TO MEASURE AN AI SYSTEM'S PERFORMANCE ON A SPECIFIC CAPABILITY, SUCH AS CODING, REASONING, FACTUAL RECALL, MATHEMATICAL PROBLEM-SOLVING, OR SCIENTIFIC ANALYSIS.
ALIGNMENT	THE CHALLENGE OF ENSURING THAT AI SYSTEMS BEHAVE IN WAYS CONSISTENT WITH HUMAN INTENTIONS, VALUES, AND CONSTRAINTS.
DUAL-USE CAPABILITY	A CAPABILITY THAT CAN SUPPORT BENEFICIAL OR HARMFUL USES DEPENDING ON CONTEXT, SAFEGUARDS, AND USER INTENT.
JAGGED TECHNOLOGICAL FRONTIER	A SITUATION WHERE AN AI SYSTEM CAN SUCCEED ON ONE TASK BUT FAIL ON ANOTHER THAT SEEMS OF EQUIVALENT DIFFICULTY TO HUMANS.
CAPABILITY–RELIABILITY GAP	A SITUATION WHERE AN AI MODEL HAS THE CAPABILITY TO PERFORM A TASK, BUT NOT DO IT CONSISTENTLY ENOUGH FOR DEPLOYMENT.
SOCIO-TECHNICAL SYSTEM	A SYSTEM MADE UP OF BOTH TECHNICAL COMPONENTS AND THE PEOPLE, INSTITUTIONS, WORKFLOWS, INCENTIVES, AND RULES THAT SHAPE HOW THOSE COMPONENTS ARE USED.

The same systems are also dual-use. Systems that assist biological research may also lower barriers to harmful experimentation.[10] Tools that strengthen cyber defense may also improve offensive capabilities.[11] Policymakers therefore need to think not only about what AI systems can do, but who can access them, how reliably they operate, and whether they advantage attackers or defenders in practice.

Despite significant and rapid progress, AI capabilities still vary unevenly across tasks. Current systems often display this "jagged technological frontier", where they perform extremely well in some areas while failing unexpectedly in others that appear similarly difficult, or even easier, to humans.[12]

This jaggedness is closely related to the "capability–reliability gap" shown by current systems: they can produce striking results on benchmarks—standardized tests used to measure model performance—while remaining inconsistent or brittle in deployment.[13] A model may solve a difficult coding problem once and fail on a closely related one the next time. An agent may appear competent on a benchmark while behaving unpredictably across runs and failing in ways that are difficult to anticipate. Rising accuracy on standard evaluations can obscure more limited progress on metrics such as consistency, robustness, predictability, and safety.[14] This means that "the model can do X" often does not mean that "the system can be relied upon to do X in the real world within acceptable error bounds."

Evaluating AI systems

In spite of these reliability issues, AI systems are increasingly being integrated into high-stakes workflows and operations. As this adoption accelerates, ensuring that the systems perform as intended becomes more important. This challenge falls within the field of testing, evaluation, verification, and validation (TEVV).[15]

Testing examines system behavior under specified conditions. Evaluation assesses capability, risk, or value against defined criteria. Verification asks whether a system was built according to its design requirements: in other words, did we build the system right? Validation, in turn, asks whether the system is fit for its intended real-world purpose: in other words, did we build the right system?

AI TEVV remains uneven and underdeveloped for general-purpose frontier systems, with major open challenges. One central problem is construct validity: whether an evaluation actually measures the capability it is supposed to measure.[16] A benchmark can be technically precise while still measuring only a narrower proxy for a broader capability, and strong performance may not generalize to operational settings. For example, a model that performs well on the LSAT has not thereby demonstrated the ability to practice law. A model that scores highly on virology questions has not necessarily shown that it can help a user carry out harmful biological work.

10 Toby Webster, Richard Moulange, Barbara Del Castello, James Walker, Sana Zakaria, and Cassidy Nelson, Global Risk Index for AI-Enabled Biological Tools: Summary Assessment & Methods Report (Cambridge, UK: Centre for Long-Term Resilience and RAND Europe, 2025), https://www.rand.org/randeurope/research/projects/2024/ai-risk-index.html

11 Nicholas Carlini, Keane Lucas, Evyatar Ben Asher, Newton Cheng, Hasnain Lakhani, David Forsythe, and Kyla Guru, "Evaluating and Mitigating the Growing Risk of LLM-Discovered 0-Days," February 5, 2026. Accessed May 9, 2026, https://red.anthropic.com/2026/zero-days/

12 For sources, please see Fabrizio Dell'Acqua et al., "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality," Organization Science 37, no. 2 (2026): pp. 403 - 423, https://pubsonline.informs.org/doi/10.1287/orsc.2025.21838 and Helen Toner, "Taking Jaggedness Seriously," Rising Tide (Substack), November 24, 2025, https://helentoner.substack.com/p/taking-jaggedness-seriously

13 Stephen Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan, "Towards a Science of AI Agent Reliability,"arXiv preprint arXiv:2602.16666, revised February 23, 2026, https://arxiv.org/abs/2602.16666.

14 Stephen Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan, "Towards a Science of AI Agent Reliability,"arXiv preprint arXiv:2602.16666, revised February 23, 2026, https://arxiv.org/abs/2602.16666.

15 National Institute of Standards and Technology, "AI Test, Evaluation, Validation, and Verification (TEVV)," accessed May 9, 2026, https://www.nist.gov/ai-test-evaluation-validation-and-verification-tevv

16 Andrew M. Bean et al., "Measuring What Matters: Construct Validity in Large Language Model Benchmarks," arXiv, November 3, 2025, https://doi.org/10.48550/arXiv.2511.04703; Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, Alex Hanna, and Amandalynne Paullada, "AI and the Everything in the Whole Wide World Benchmark," arXiv preprint arXiv:2111.15366, November 26, 2021, https://arxiv.org/abs/2111.15366.

Even so, influential benchmarks are often treated as broad indicators of progress toward general intelligence or dangerous capabilities. High scores can therefore mislead policymakers into viewing systems as more general, robust, or dangerous than the underlying evidence supports. Low scores can also obscure risks that emerge only when systems are connected to tools, embedded in workflows, or used by motivated actors.

AI capabilities are also moving faster than evaluation methods are maturing. By the time evaluations can provide strong evidence about a model's capabilities, limitations, or real-world impacts, the frontier may already have shifted.[17] At the same time, the object of evaluation keeps changing. Evaluating a base model is different from evaluating a tool-using system, an AI agent, or a model embedded in a complex sociotechnical workflow. Each adds new sources of uncertainty: longer time horizons, more interaction with external systems, greater dependence on human users, and more opportunities for failures that do not appear in static benchmarks.[18] There is even a growing phenomenon of "evaluation awareness", where AI models alter their behavior during evaluation compared to real world performance.[19] Together, these challenges make it increasingly difficult to turn a messy, evolving system into a single capability score, and it is hard to know where AI systems will succeed and fail when they encounter complex real-world tasks.

Three views of AI's trajectory

Most policy disagreements around AI can be organized into three broad camps. These are not rigid groups, and many people hold views that span multiple perspectives. Still, the categories are useful because they reveal the main assumptions structuring contemporary policy debates. The camps are separated by their beliefs about how capable AI systems will be, but we also see a deeper difference concerning agency: who or what does each camp believe to be the relevant actors?

MIRAGE: AI AS CON, BUBBLE, OR CHARISMATIC TECHNOLOGY

In this camp, the relevant actors are humans and institutions, including the developers and deployers of AI tools and systems. AI itself is not the mover, and indeed "AI" may be seen as purely a branding term. The central story is about businesses, investors, deployers, commentators, policymakers, and users who project extraordinary significance onto a still-limited technology. In this camp, AI could be described as a "charismatic technology," one that has acquired an unusual grip over public imagination by attracting utopian visions of progress, salvation, inevitability, and moral mission that far exceed its demonstrated capabilities.[20] From this standpoint, the technical system becomes less important than the devotion it elicits from people.

People within this camp see current systems as fundamentally limited, and expect those limitations to persist. The apparent generality of AI is an illusion created by benchmark design, selective demos, or the human tendency to anthropomorphize language output. Benchmarks can be gamed, and eye-catching performance in narrow settings

17 Yoshua Bengio et al., International AI Safety Report 2026, DSIT 2026/001, February 3, 2026, https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026; Stephen Casper, David Krueger, and Dylan Hadfield-Menell, "Pitfalls of Evidence-Based AI Policy," arXiv, last revised September 15, 2025, https://doi.org/10.48550/arXiv.2502.09618

18 Sayash Kapoor et al., "Open-World Evaluations for Measuring Frontier AI Capabilities," AI as Normal Technology, April 16, 2026, https://www.normaltech.ai/p/open-world-evaluations-for-measuring.

19 Sambhav Maheshwari and Joe O'Brien, "Evaluation Awareness: Why Frontier AI Models Are Getting Harder to Test," Institute for AI Policy and Strategy, March 31, 2026, https://www.iaps.ai/research/evaluation-awareness-why-frontier-ai-models-are-getting-harder-to-test; Marcus Williams, Cameron Raymond, and Micah Carroll, "Sidestepping Evaluation Awareness and Anticipating Misalignment with Production Evaluations," OpenAI Alignment Blog, December 18, 2025, https://alignment.openai.com/prod-evals/; Joe Needham et al., "Large Language Models Often Know When They Are Being Evaluated," arXiv, last revised July 16, 2025, https://doi.org/10.48550/arXiv.2505.23836

20 Morgan G. Ames, "Charismatic Technology," in Proceedings of CC 2015, the 5th Decennial Conference on Critical Computing, Aarhus, Denmark, August 2015 (ACM Press, 2015), 109–20, https://www.morganya.org/research/Ames-charisma-aarhus.pdf; Ritwik Gupta and Andrew W. Reddie, "The LLM Mirage: Economic Interests and the Subversion of Weaponization Controls," arXiv, submitted January 8, 2026, https://doi.org/10.48550/arXiv.2601.05307.

is routinely mistaken for robust understanding or autonomy. AI companies may be running into diminishing returns or economic constraints, and are unlikely to turn profitable. The grandest claims about "artificial general intelligence" or "superintelligence" are therefore read less as scientific forecasts than as ideological or commercial projects to continue the flow of investment.

This camp interprets AI through the history of technologies whose promises ran ahead of their demonstrated value. One example is AI itself, and its history of AI "summers" and "winters": repeated episodes in which grand claims outran real capability and periods of enthusiasm were followed by disappointment and funding retrenchment.[21] Another is cryptocurrencies, where speculative capital, evangelizing rhetoric, and promises of civilizational transformation often dwarf actual socially valuable deployment. Within this camp, AI may be real and useful in some specific domains while being far from what its most enthusiastic advocates claim.

The opportunities under this trajectory are far narrower than the AI boosters suggest, but they are still real. Technologies labeled "AI" can still be useful tools. Better procurement, careful use-case selection, and greater skepticism could still provide benefits without requiring society to reorganize around metaphysical claims about "machine minds." In this view, the main policy gains would come from resisting utopian thinking, and focusing on where there is strong evidence of specific tools already working.

A significant challenge for policymakers from this perspective is capital misallocation: an AI bubble can distort investment and produce political pressure to deploy systems before they are ready. Another risk is policymaker distraction: the government may focus on future risks and benefits while neglecting the ordinary but damaging problems of fraud, labor exploitation, poor procurement, environmental cost, bias, and systems that simply don't work. Limited but widely adopted technology can still cause real harm.

DIFFUSION: AI AS NORMAL TECHNOLOGY

In the next camp, AI is neither a con nor a new superintelligent species. It is a powerful but ultimately "normal" technology. 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, in a similar fashion to how society has restructured itself around previous general-purpose technologies. 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."[22]

AI could still be highly important, even transformative, but its effects will likely arrive through diffusion, adaptation, and institutional redesign over decades rather than through a sudden transition. The historical analogies here are general-purpose technologies such as steam and electricity. Electricity mattered enormously, but its productivity effects depended on the reorganization of factories and complementary investments. On this view, the major policy questions concern adoption, training, interoperability, regulation, organizational learning, and how power shifts when some actors integrate these systems more effectively than others.

This camp also draws on present evidence. Jaggedness and the capability–reliability gap described in the previous section suggest that current systems are better used in collaboration with human systems rather than as autonomous replacements for them. That is why this view emphasizes sociotechnical systems rather than standalone AI models.

The opportunities under this trajectory are substantial. AI could accelerate scientific work, improve planning and analysis, support logistics, widen access to technical assistance, and increase productivity in some sectors without immediately producing societal disruption. Governments could use these tools for better service delivery and administrative capacity. In national security settings, AI could improve intelligence, logistics, and decision support

21 Daniel Crevier, AI: The Tumultuous History of the Search for Artificial Intelligence (New York: Basic Books, 1993).
22 Arvind Narayanan and Sayash Kapoor, "AI as Normal Technology," Knight First Amendment Institute, April 15, 2025, https://knightcolumbia.org/content/ai-as-normal-technology.

while still leaving agency for core oversight with humans and organizations. But all of this only happens with careful integration and human supervision.

The risks under this trajectory are less cinematic than loss of control to "superintelligence", and are centered on the interaction of humans with AI and the incorporation of AI into safety-critical sectors. First, there is the risk of brittle integration: models that look useful in pilot settings can fail when dealing with real-world messiness. Second, there is automation bias: people may over-trust systems that are good enough to be seductive but not good enough to be relied on. Third, there is uneven diffusion. Some organizations and states may integrate AI faster and better than others, creating distributional and geopolitical effects even if the underlying technology remains "normal."[23] Fourth, there is the diffusion of capabilities to a broader set of less sophisticated actors. AI may not only amplify what the best-resourced actors can do; it may also widen what actors with limited technical abilities can achieve. That could matter greatly for areas like cybersecurity,[24] as well as biosecurity and other global risk domains.

In this camp, the main policy challenge is governing a capable but brittle dual-use technology as it diffuses through institutions that already have weaknesses and are often slow to adapt. This is a more mundane challenge than rogue superintelligence, but not a minor one. AI may still produce a transition comparable in scale to previous industrial revolutions, even if that transition is uneven, institutional, and gradual.

AUTONOMOUS POWER: AI ON A PATH TO ARTIFICIAL GENERAL INTELLIGENCE (AGI) AND ARTIFICIAL SUPER INTELLIGENCE (ASI)

In the last camp, the relevant actor shifts beyond humans alone. Humans still matter enormously, at least in the short term, with AI developers and states as crucial actors. But the policy question expands from what humans can do with AI to what increasingly capable and autonomous AI systems may do themselves. In this view, AI may become an actor in its own right, or at least enough of one that its abilities and motives can no longer be treated as directly flowing from human intent.

The core claims of this camp are that: 1) current systems may be on a trajectory toward qualitatively new capabilities, well beyond those of humans; 2) discontinuities or very steep gradients in real-world impacts are plausible in years, not decades; and 3) sufficiently powerful systems may be difficult to control even if they are initially built by well-intentioned actors.[25] This camp will often focus on "artificial general intelligence" (AGI, often meaning AI as capable as humans on cognitive tasks) and "artificial superintelligence" (ASI, often meaning AI far more capable than humans on cognitive tasks) as key milestones, although these terms have increasingly attracted a range of conflicting definitions.[26]

What evidence does this perspective rely on? One example is the pace of capability progress in main domains,[27] particularly software engineering.[28] This progress could lead to the increasing automation of the research-and-

23 Jeffrey Ding, Technology and the Rise of Great Powers: How Diffusion Shapes Economic Competition (Princeton, NJ: Princeton University Press, 2024), https://press.princeton.edu/books/paperback/9780691260341/technology-and-the-rise-of-great-powers

24 Nicholas Carlini et al., "Evaluating and Mitigating the Growing Risk of LLM-Discovered 0-Days," Anthropic, February 5, 2026, https://red.anthropic.com/2026/zero-days/

25 Dan Hendrycks, Mantas Mazeika, and Thomas Woodside, "An Overview of Catastrophic AI Risks," arXiv, last revised October 9, 2023, https://doi.org/10.48550/arXiv.2306.12001; Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean, "AI 2027," April 3, 2025, https://ai-2027.com/; Center for AI Safety, "Statement on AI Risk," May 30, 2023, https://aistatement.com/; Yoshua Bengio, "How Rogue AIs May Arise," Yoshua Bengio, May 22, 2023, https://yoshuabengio.org/en/blog/how-rogue-ais-may-arise; Future of Life Institute, "Statement on Superintelligence," March 27, 2026, https://superintelligence-statement.org/; Anthony Aguirre, "Keep the Future Human," Keep the Future Human, March 5, 2025, https://keepthefuturehuman.ai/.

26 Dan Hendrycks et al., "A Definition of AGI," arXiv, last revised December 3, 2025, https://doi.org/10.48550/arXiv.2510.18212; Helen Toner, "The Term 'AGI' Is Almost Useless at This Point," Rising Tide, April 6, 2026, https://helentoner.substack.com/p/the-term-agi-is-almost-useless-at; Meredith Ringel Morris et al., "Levels of AGI for Operationalizing Progress on the Path to AGI," arXiv, last revised September 24, 2025, https://doi.org/10.48550/arXiv.2311.02462.

27 Epoch AI, "Epoch Capabilities Index," accessed April 29, 2026, https://epoch.ai/eci.

28 Thomas Kwa et al., "Measuring AI Ability to Complete Long Tasks," METR, March 19, 2025, https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/.

development process that produces new AI systems, which would in turn speed up the rate of AI progress, resulting in an "intelligence explosion."[29] A second source of support is more conceptual: philosophical and decision-theoretic reasoning about what highly capable goal-directed agents may tend to do. On this view, a system pursuing almost any sufficiently ambitious objective may have instrumental reasons to acquire resources, preserve its own operation, avoid shutdown, shape its environment, and gain power over the processes that determine whether it succeeds.[30]

The reference classes for this view differ sharply from those in the other two camps. Rather than comparing AI to electricity, steam power, or earlier speculative technologies, proponents often draw on analogies involving biological evolution, unfamiliar forms of intelligence, or impactful biological events like pandemics. These analogies are imperfect and often sit outside conventional policy discourse. Still, they reflect the core premise of this worldview: that AI could create sharp asymmetries in cognition, autonomy, and power that make comparisons to earlier technologies inadequate.

The opportunities under this trajectory are enormous. If one takes the strongest versions of this view seriously, AI could radically accelerate scientific discovery, automate large portions of technical labor, improve medicine and biology, and produce levels of abundance and problem-solving capacity that are otherwise unimaginable. This camp contains some of the most ambitious visions of human flourishing through AI, as well as visions for "transhumanist" human-AI mergers.

But this view also contains the sharpest version of the control problem. This camp may still be concerned about the misuse of AI by malicious humans, accidents in safety-critical systems, or systems with limited autonomy evading human oversight, similar to concerns that could be found within the "normal technology" camp. However, these concerns are far less pressing than the idea that highly capable systems may pursue power acquisition and self-preservation in ways that present existential risks to humanity. Within this worldview, even if a chance of extinction is low or highly uncertain, that could still be enough to make managing catastrophic AI risks a top priority.

For policymakers, this camp presents the challenge of balancing immense opportunity with huge risks, all on a foundation of extreme uncertainty. The strongest upside claims and strongest downside claims stem from the same premise: that sufficiently capable AI could become a very general force multiplier and an extremely capable autonomous actor. This camp, therefore, often pairs broad optimism about AI's impact on science, engineering, and economic growth from applications of powerful AI with serious concern about control, misuse, and power concentration.[31]

Why the camps diverge

These three camps have numerous disagreements: how to evaluate current AI systems, what policy questions are worthy of attention, and what future AI systems may look like. They also disagree more fundamentally about what object they are analyzing, and how to see the relevant actors. Six disagreements are especially important.

The first is agency. In the mirage camp, humans are the only meaningful actors. AI matters because people promote it, invest in it, deploy it badly, and get seduced by its "charisma." In the normal technology camp, humans remain the actors, but the relevant unit is often the human–AI system or the institution that embeds AI into its operations.

29 Alan Chan et al., "Measuring AI R&D Automation," arXiv, last revised March 6, 2026, https://doi.org/10.48550/arXiv.2603.03992; Severin Field, Raymond Douglas, and David Krueger, "AI Researchers' Views on Automating AI R&D and Intelligence Explosions," arXiv, last revised March 5, 2026, https://doi.org/10.48550/arXiv.2603.03338.

30 Joseph Carlsmith, "Is Power-Seeking AI an Existential Risk?," arXiv, last revised August 13, 2024, https://doi.org/10.48550/arXiv.2206.13353.

31 Dario Amodei, "Machines of Loving Grace: How AI Could Transform the World for the Better," October 2024, https://www.darioamodei.com/essay/machines-of-loving-grace; Dario Amodei, "The Adolescence of Technology: Confronting and Overcoming the Risks of Powerful AI," January 2026, https://www.darioamodei.com/essay/the-adolescence-of-technology; Yoshua Bengio, "Reasoning Through Arguments Against Taking AI Safety Seriously," YoshuaBengio.org, July 9, 2024, https://yoshuabengio.org/en/blog/reasoning-through-arguments-against-taking-ai-safety-seriously.

In the autonomous power camp, the system itself increasingly becomes part of the actor model. This distinction determines what kinds of risk one sees as primary. If AI is not an actor, then AI control problems look overdrawn. If AI may become an actor, then ordinary misuse frameworks start to look incomplete.

Second, the camps disagree about units of analysis. Some people focus on specific AI models. Others focus on deployed systems and their interactions with humans. Still others focus on individuals, organizations, markets, or states. That is one reason the debates can be so confusing: the participants are often not arguing about the same thing. One person is looking at benchmark curves. Another is looking at public-sector procurement failures. Another is looking at long-run control problems of hypothetical systems. All may be making valid observations, but about different objects.

Third, they disagree about reference classes. The power of reference class selection is particularly strong when we are trying to speculate about an uncertain future in the presence of incomplete evidence. To treat AI as either like cryptocurrencies, electricity, or an alien species means the chosen class imports profoundly different expectations about AI's trajectory. Many policy disagreements are really downstream from different analogies and the historical lessons commentators believe we can learn from those analogies.

Fourth, they disagree about how to manage evidence. Some analysts treat benchmark trends and model scaling as the relevant evidence, using these to project to future capabilities and impacts. Others put more weight on real-world deployment failures, labor market data, or the slow pace of adoption in high-stakes institutions. Others again think that waiting for direct empirical evidence of the most extreme risks is misguided because those risks may materialize before a robust evidence base can form—in the absence of such direct evidence, they will often rely on thought experiments and extrapolations.[32]

Fifth, they disagree about pace. Is the binding constraint innovation at the frontier of AI models or diffusion through the economy? Will benchmark progress translate rapidly into real-world transformation, will institutional friction dominate, or will advances stagnate? Will there be a steep takeoff in some capabilities but slow adoption elsewhere? Will jaggedness persist over long time scales?

Sixth, they often implicitly rely on different safety and risk paradigms. One is accident-focused safety thinking, focusing on failures, brittleness, and complex-system accidents. Another is adversarial security thinking: misuse, attack pathways, prompt injection, hijacking, and the behavior of human threat actors. A third is misaligned-agent thinking: the possibility that highly capable systems themselves become difficult to control. Much of the safety literature has been shaped by accident-focused thinking, while many global-risk questions are inherently adversarial, and the ASI discourse shifts again toward AI as a novel form of threat. Different parties may use the same word—"safety"—to mean very different things, and propose to manage risk with very different tools.[33]

Policymakers must act under uncertainty

32 Stephen Casper, David Krueger, and Dylan Hadfield-Menell, "Pitfalls of Evidence-Based AI Policy," arXiv, last revised September 15, 2025, https://doi.org/10.48550/arXiv.2502.09618.

33 Dario Amodei et al., "Concrete Problems in AI Safety," arXiv, last revised July 25, 2016, https://doi.org/10.48550/arXiv.1606.06565; Inioluwa Deborah Raji and Roel Dobbe, "Concrete Problems in AI Safety, Revisited," ICLR Workshop on Machine Learning in the Real World, 2020, arXiv preprint, submitted December 18, 2023, https://doi.org/10.48550/arXiv.2401.10899; Arvind Narayanan and Sayash Kapoor, "AI Safety Is Not a Model Property," AI as Normal Technology, March 12, 2024, https://www.normaltech.ai/p/ai-safety-is-not-a-model-property; Cullen O'Keefe, "AI Safety Is Sometimes a Model Property," Jural Networks, May 2, 2024, https://juralnetworks.substack.com/p/ai-safety-is-sometimes-a-model-property; Roel I. J. Dobbe, "System Safety and Artificial Intelligence," in The Oxford Handbook of AI Governance (Oxford: Oxford University Press, 2022), 441–58, https://doi.org/10.1093/oxfordhb/9780197579329.013.67; Remco Zwetsloot and Allan Dafoe, "Thinking About Risks From AI: Accidents, Misuse and Structure," Lawfare, February 11, 2019, https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure; Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," arXiv, last revised February 22, 2018, https://doi.org/10.48550/arXiv.1802.07228.

Policymakers deal with expert disagreement all the time, but AI poses a harder version of that problem because evidence often arrives more slowly than capabilities. Former United States Secretary of the Navy Richard Danzig has argued that national security decision-makers must not only make their predictions explicit, but also prepare for those predictions to fail.[34] Policy must be designed to remain useful when forecasts turn out to be wrong, and ensuring broader perspectives in these dialogues is essential to counteract an overly narrow construction of how we frame risk.

The 2026 *International AI Safety Report*[35] captures a key AI policy challenge with the phrase "evidence dilemma": general-purpose AI capabilities evolve quickly, but it takes time to collect and assess evidence about their societal effects. AI's jagged technological frontier deepens this dilemma because uneven performance makes both capabilities and risks hard to interpret. A benchmark jump may not translate into reliable field performance, while a failed deployment may not mean the underlying capability trend is illusory. Policymakers therefore face pressure to act before evidence is settled: act too early, and they may lock in weak or misdirected interventions; wait too long, and societies may be exposed to risks that are difficult to reverse once systems have diffused.

Some policy interventions could be supported by multiple camps, though the reasons for supporting them differ. One is improving measurement and TEVV, including more context-specific and ongoing evaluation rather than a one-time "benchmark theater." Another is greater transparency about model capabilities, limits, and deployment context. A third is government capacity: agencies need staff who can use these systems, understand their limitations, and interpret claims made by developers and critics alike.

However, the camps obviously do not imply identical policies.

Policymaking, therefore, requires explicit uncertainty management. Policymakers should avoid pretending that one forecast has already been fully vindicated, but they should also avoid using uncertainty as an excuse for waiting indefinitely. They should build institutions that can gather information quickly, revise assumptions publicly, and take robust action under multiple scenarios. They should ask not just, "What do we think AI is?" but also, "What would we do if we are wrong?"

To make that discipline practical, policymakers need a way to expose the assumptions behind competing proposals. Five questions are especially useful:

- First, who or what does this proposal treat as the relevant actor: humans, human–AI systems, or AI systems themselves?
- Second, what arguments, evidence, and historical reference classes does this proposal rely on (implicitly or explicitly)?
- Third, what kind of risk or opportunity is this proposal particularly focused on?
- Fourth, what evidence would challenge or support the existence of these risks or opportunities?
- Fifth, is that evidence likely to arrive in time for policymakers to update course before the relevant risks or benefits are locked in, and can the evidence collection be accelerated?

These questions do not remove uncertainty, but do provide a more structured way to surface the driving assumptions of different policy approaches and bring fundamental disagreements into the open.

34 Richard Danzig, "Driving in the Dark: Ten Propositions About Prediction and National Security," Center for a New American Security, October 26, 2011, https://www.cnas.org/publications/reports/driving-in-the-dark-ten-propositions-about-prediction-and-national-security.

35 Yoshua Bengio et al., International AI Safety Report 2026, DSIT 2026/001, February 3, 2026, https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026.

Chapter 2. Global Risks and Artificial Intelligence

For much of the modern era, global risk has been understood as the possibility of catastrophic harm arising from distinct domains. Over the past seven decades, policymakers and experts have focused primarily on threats such as nuclear weapons, biological and chemical agents, and large-scale military conflict: risks capable of producing mass casualties, destabilizing states, or, in the most extreme cases, threatening human survival.[36]

Although these risks could have global consequences, they were generally treated as domain-specific problems managed through specialized institutions, treaties, and governance frameworks such as the Nuclear Nonproliferation Treaty, Chemical Weapons Convention, and the Open Skies Treaty. The central challenge was preventing intentional misuse or catastrophic failure in systems that were comparatively well-characterized and physically bounded.

AI complicates this picture significantly. The "normal technology" and "autonomous power" perspectives described in Chapter 1 locate uncertainty in different places.

Where do these two perspectives converge? Each perspective locates uncertainty in the risks at different points in the safety, security, and innovation process. [37] From a "normal technology" perspective, uncertainty exists when AI is used or integrated into increasingly complex and interconnected systems. This means discerning and enforcing "red lines" of use and adoption might bound these risks.[38] From the "autonomous power" perspective, our opportunity to anticipate and address governance challenges proactively may already be closing, with companies and governments left to struggle to understand and contain rapidly evolving and opaque AI systems.[39]

Presently, accelerating AI development, wider access to technical capabilities, and the integration of AI into already complex systems are compressing decision timelines and exposing new vulnerabilities in existing infrastructures. As a result, global risks increasingly cut across domains rather than remaining confined within them as they intersect further with AI. Further, this implies that historic governance frameworks, which were designed for a slower, more compartmentalized world, are now under growing strain.[40]

Global risks, therefore, are increasingly shaped not by any single technology or actor, but by the interaction between rapidly advancing capabilities and institutions that were not designed to originally absorb them.[41] Much as the Internet transformed communication, commerce, and security faster than governance systems could adapt, AI is now accelerating change across biotechnology, cyber operations, and military decision-support systems while compressing timelines for detection, decision-making and response.[42]

Geopolitical competition and commercial pressures further complicate these risks by weakening incentives to slow deployment or invest in shared risk reduction. Addressing these challenges does not require halting

36 John P. Caves and Seth W. Carus, The Future of Weapons of Mass Destruction (Washington, DC: Center for the Study of Weapons of Mass Destruction, National Defense University, 2012), Occasional Paper No. 10, https://ndupress.ndu.edu/Portals/97/Documents/Publications/Occasional%20Papers/10_Future%20of%20WMD.pdf

37 Seán Ó hÉigeartaigh, "Stopping the Clock on Catastrophic AI Risk," Bulletin of the Atomic Scientists Vol. 81, No. 6 (2025): pp. 462 - 467, https://www.tandfonline.com/doi/epdf/10.1080/00963402.2025.2586972?needAccess=true

38 For sources, please see International Dialogues on AI Safety, "IDAIS-Beijing, 2024," International Dialogues on AI Safety, accessed May 9, 2026, https://idais.ai/dialogue/idais-beijing/ and "AI Red Lines: The Opportunities and Challenges of Setting Limits," World Economic Forum, March 11, 2025, https://www.weforum.org/stories/2025/03/ai-red-lines-uses-behaviours/

39 Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean, AI 2027, AI Futures Project, April 3, 2025, https://ai-2027.com/

40 For sources, please view the series of roundtable memos we have developed through expert input and dialogue through this project. These may be found here: https://fas.org/initiative/ai-x-global-risk-nexus-project/

41 World Economic Forum, The Global Risks Report 2026, 21st ed. Geneva: World Economic Forum, January 14, 2026, https://reports.weforum.org/docs/WEF_Global_Risks_Report_2026.pdf

42 For sources, please view the series of roundtable memos we have developed through expert input and dialogue through this project. These may be found here: https://fas.org/initiative/ai-x-global-risk-nexus-project/

innovation: rather, it requires coordinated approaches that align safety, security, and resilience with technical understanding and effective governance.[43]

This chapter introduces a shared framework for considering global risk through three interacting components: threat, vulnerability, and consequence. It then examines how AI reshapes these dynamics and draws on risk mitigation and high reliability frameworks to explain how increasingly complex systems generate and manage risk across local, national, and global scales. Together, these concepts establish the foundation for the chapters that follow.

Threat, Vulnerability, and Consequence - A Working Model of Risk

One way to think about global risk is through the interaction between threat, vulnerability, and consequence (TVC). This framing is often presented as a simple heuristic:[44]

Risk = Threat x Vulnerability x Consequence

In this equation, **THREAT** refers to the actors, intentions, and capabilities that create pathways for harm. **VULNERABILITY** refers to the weaknesses or points of failure within systems that allow threats to manifest. **CONSEQUENCE** refers to the outcomes that result when threats exploit vulnerabilities. These can range from localized disruptions to large-scale systemic effects, and are shaped not only by the initial event, but by how it propagates across interconnected systems and how actors respond.

This heuristic showcases three key points. First, risk is not reducible to a single bad actor, dangerous technology, or catastrophic outcome in isolation. Risk emerges from the interaction between multiple factors that increase or decrease the likelihood and severity of harm. Polio demonstrates this dynamic well. Historically, polio posed high threat, vulnerability, and consequence due to widespread transmission, population susceptibility, and severe health outcomes. Global vaccination campaigns sharply reduced both vulnerability and effective threat, dramatically lowering overall risk without eliminating the virus entirely.[45]

Second, this framing highlights how risk changes over time as threats, vulnerabilities, and consequences evolve together. Cybersecurity provides a clear example. The rise of ransomware reflects not only more capable malicious actors, but also persistent vulnerabilities across public and private digital infrastructure.[46] Over time, offensive tools, exploit marketplaces, and technical capabilities have become more widely accessible, while many critical systems remain insufficiently hardened.[47] Attacks such as Petya and NotPetya further demonstrate how cyber risks

43 Ibid.

44 This is a typical framing for national security risk assessments in areas such as the U.S. Intelligence Community (IC) and the U.S. Department of Homeland Security. Elements of this are also present in more technologically focused areas such as the U.S. National Institute of Standards and Technology. For sources, please see U.S. Department of Homeland Security, Risk Steering Committee, DHS Risk Lexicon (Washington, DC: U.S. Department of Homeland Security, September 2010), https://www.cisa.gov/sites/default/files/publications/dhs-risk-lexicon-2010_0.pdf; Office of the Director of National Intelligence, National Counterintelligence and Security Center, Framework for Assessing Risk (Washington, DC: Office of the Director of National Intelligence, 2021), https://www.dni.gov/files/NCSC/documents/supplychain/Framework_for_Assessing_Risks_-_FINAL_Doc.pdf; and Joint Task Force Transformative Initiative, Guide for Conducting Risk Assessments, NIST Special Publication 800-30 Rev. 1 (Gaithersburg, MD: National Institute of Standards and Technology, September 2012), https://csrc.nist.gov/pubs/sp/800/30/r1/final.

45 World Health Organization, WHO Global Action Plan for Poliovirus Containment, 4th edition (GAPIV) (Geneva: World Health Organization, 2022), https://polioeradication.org/wp-content/uploads/2022/07/WHO-Global-Action-Plan-for-Poliovirus-Containment-GAPIV.pdf. Accessed March 29, 2026.

46 Stuart E. Madnick, The Continued Threat to Personal Data: Key Factors Behind the 2023 Increase (Cambridge, MA: Massachusetts Institute of Technology, 2023), https://www.apple.com/newsroom/pdfs/The-Continued-Threat-to-Personal-Data-Key-Factors-Behind-the-2023-Increase.pdf

47 UK National Cyber Security Centre. "Mitigating Malware and Ransomware Attacks." NCSC.GOV.UK. February 13, 2020. https://www.ncsc.gov.uk/guidance/mitigating-malware-and-ransomware-attacks

can escalate in both scale and consequence.[48] NotPetya, in particular, spread automatically across interconnected systems and caused widespread disruption far beyond its original target.[49]

Finally, this framing highlights that assessing risk is not sufficient: it must be actively shaped by targeting its underlying components. By understanding whether threat, vulnerability, or consequence is most tractable in a given context, policymakers and practitioners can intervene to most effectively *reduce overall risk*. Nuclear arms control during the Cold War illustrates this principle. Confidence-building measures, hotlines, and early-warning protocols reduced vulnerability to misperception and inadvertent escalation, while treaties such as New START sought to constrain potential consequences by limiting arsenal size. These efforts did not eliminate nuclear risk, but they demonstrate how even high-consequence technologies can be managed through interventions across multiple dimensions of risk.[50],[51]

At the same time, this framework should not be mistaken for a law of nature. Experts have warned that threat, vulnerability, and consequences do not always behave like straightforward, independent variables. For example, threat actors adapt. Further, vulnerabilities may shift in response to defense measures, and consequences may "proliferate" and drive future behavior, incentives, and escalatory actions. Therefore, this framing is often best understood as a heuristic, and not a literal equation. This caveat becomes important as we consider how today's high-risk system architectures are characterized by complexity, tight coupling, opacity, and rapid interaction effects. In such settings, the line between threat and vulnerability can blur, and consequences can propagate in ways that are nonlinear and difficult to forecast.[52]

HOW AI TRANSFORMS RISK: FROM DOMAIN-SPECIFIC THREATS TO SYSTEM-LEVEL CHALLENGES

Taken together, these examples demonstrate both the utility and flexibility of this TVC framework. Risk can often be reduced not by eliminating threats entirely, but by reshaping vulnerabilities, limiting consequences, or altering how these factors interact over time. At the same time, these examples largely reflect risks that remain tractable within identifiable domains. AI complicates this picture in three key ways.

First, as a general-purpose and cross-domain technology, AI interacts with each component of the framework simultaneously. It both expands the range and capability of threats by lowering barriers to entry or amplifying existing actors, as well as potentially becoming a threat actor itself as AI systems become increasingly autonomous. It can also introduce new vulnerabilities by increasing system complexity, opacity, and interdependence.[53]

48 Cybersecurity and Infrastructure Security Agency, "Petya Ransomware," CISA Alerts, July 1, 2017, https://www.cisa.gov/news-events/alerts/2017/07/01/petya-ransomware

49 Andy Greenberg, Sandworm: A New Era of Cyberwar and the Hunt for the Kremlin's Most Dangerous Hackers. New York: Doubleday, 2019.

50 Eric Schlosser, Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety (New York: Penguin Books, 2014).

51 While much has been written on this topic, and often with different interpretations of the case studies and the effectiveness of these measures, the following sources provide a base grounding for the discussion above. They include Michael Krepon, Nuclear Risk Reduction: Moving Beyond the Cold War (Washington, DC: Henry L. Stimson Center, 2004), https://www.stimson.org/wp-content/files/NRRMKrepon.pdf; Nikolai Sokov, Sahil V. Shah, David Santoro, and Miles Pomper, Reimagining Risk Reduction: Adapting Cold War Tools to Manage 21st Century Strategic Instability (Vienna: Vienna Center for Disarmament and Non-Proliferation, February 2025), https://vcdnp.org/wp-content/uploads/2025/02/VCDNP_Reimagining-Risk-Reduction_web.pdf; and Tytti Erästö and Wilfred Wan, "Risk Reduction is Urgently Needed Amid Rising Tensions in Northern Europe," Stockholm International Peace Research Institute (SIPRI), December 16, 2025, https://vcdnp.org/wp-content/uploads/2025/02/VCDNP_Reimagining-Risk-Reduction_web.pdf.

52 For sources, please see Louis Anthony Tony Cox, "Some Limitations of 'Risk = Threat x Vulnerability x Consequence' for Risk Analysis of Terrorist Attacks," Risk Analysis, Vol. 28, No. 6 (December, 2008): pp. 1749 - 1761, https://pubmed.ncbi.nlm.nih.gov/19000071/; National Research Council, "Chapter 3: Challenges to Risk Analysis for Homeland Security," in Review of the Department of Homeland Security's Approach to Risk Analysis (Washington, DC: National Academies Press, 2010), https://www.nationalacademies.org/read/12972/chapter/5; and Charles Perrow, Normal Accidents: Living with High-Risk Technologies (New York: Basic Books, 1984).

53 Johns Hopkins School of Advanced International Studies, "The Role of AI in Reducing the Risk of Weapons of Mass Destruction," SAIS News and Press, August 1, 2025, https://sais.jhu.edu/news-press/event-recap/role-ai-reducing-risk-weapons-mass-destruction. Accessed March 30, 2026.

Second, capabilities developed in one domain may be applied in another. Systems that were once separable and siloed have become increasingly integrated with AI tools and capabilities. This leads to outcomes like the compression of decision-making timelines, as well as a reduction of opportunities for actions like verification, coordination, and intervention.[54] As a result, the assumption that threat, vulnerability, and consequence can be separately considered, analyzed, and reduced becomes more difficult. Risk is not just about these individual variables, but increasingly about how they evolve together within complex systems.[55]

Third, AI alters consequences by accelerating the speed at which events unfold, as well as expand their scale or even enable effects that propagate across domains that were previously more loosely intertwined. AI-integrated sensors, autonomous systems, and increasingly networked decision-support architectures compress timelines, particularly during crisis settings, for mission-critical activities like interpretation, deliberation, and response. This is especially concerning in domains such as military command-and-control, cyber operations, and protecting critical infrastructure.[56]

AI as a Source of Threat: Capability, Intent, and Control

Chapter 1 outlined three broad ways of understanding AI: as an overhyped and potentially limited technology, as a powerful but ultimately normal technology, and as a pathway toward increasingly autonomous systems that may become difficult to control. Up to this point, this chapter has largely approached AI through the normal-technology perspective, with an emphasis on how institutions integrate and govern AI systems over time.[57]

Under this normal technology view, the primary source of threat remains the human actor, potentially amplified by AI capabilities. In the autonomous power perspective, however, the actor model begins to change. Concerns shift from how humans use AI to how increasingly capable systems behave, interact, and evolve with reduced human oversight.

One driver of risk within this view is the emergence of novel system behaviors. As AI capabilities scale and are integrated across domains, systems may produce outcomes that were never explicitly programmed: rather, they arise from interactions between components, models, and human operators. These behaviors are difficult to predict because they do not stem from any single model, dataset, or decision-point. Further, some of these behaviors, such as evaluation awareness where models recognize they are being tested, complicate both safety evaluation and governance.[58]

The opaque nature of these systems is also a challenge. Many advanced AI systems, particularly those based on large-scale machine learning architectures, operate in ways that are difficult to interpret for even their developers.[59] This creates direct challenges for risk assessment and governance. When the internal logic of a system is poorly understood, it becomes harder to anticipate how it will behave under stress, fail in unfamiliar environments, or

54 Jeffrey Ding, "Machine Failing: How Systems Acquisition and Software Development Flaws Contribute to Military Accidents," Texas National Security Review, Vol. 8, No. 1 (Winter 2024/2025), https://tnsr.org/2024/10/machine-failing-how-systems-acquisition-and-software-development-flaws-contribute-to-military-accidents/. Accessed March 30, 2026.

55 Andrés Ilcic, Miguel Fuentes, and Diego Lawler, "Artificial Intelligence, Complexity, and Systemic Resilience in Global Governance," Frontiers in Artificial Intelligence (2025), https://pmc.ncbi.nlm.nih.gov/articles/PMC12171231/. Accessed March 30, 2026.

56 Harold Trinkunas and Herbert S. Lin, "Introduction: Emerging Technologies and the Future of Strategic Stability," Texas National Security Review, February 25, 2026, https://tnsr.org/roundtable/emerging-technologies-and-the-future-of-strategic-stability/.

57 Arvind Narayanan and Sayash Kapoor, "A Guide to Understanding AI as Normal Technology," AI as Normal Technology (Substack), September 9, 2025, https://www.normaltech.ai/p/a-guide-to-understanding-ai-as-normal?open=false#%C2%A7normal-doesnt-mean-mundane-or-predictable

58 Thomas Woodside, "Emergent Abilities in Large Language Models: An Explainer," Center for Security and Emerging Technology (CSET), Georgetown University, April 16, 2024, https://cset.georgetown.edu/article/emergent-abilities-in-large-language-models-an-explainer/. Accessed March 30, 2026.

59 The Lancet Digital Health, "Large Language Models and Misinformation," The Lancet Digital Health, published online January 2026, https://www.thelancet.com/journals/landig/article/PIIS2589-7500(25)00157-8/fulltext

interact with other systems.[60] Therefore, institutions may become increasingly reliant on post-hoc detection and response rather than proactive mitigation.[61]

The introduction of multi-agent systems further complicates this picture. As multiple AI systems are deployed in tandem in areas like finance, cyber operations, military contexts, or scientific discovery, they may interact with one another in ways that are not centrally coordinated or fully visible to human operators. These interactions can create outcomes that evolve at speeds and levels of complexity beyond human oversight. In other words, creating interlinked systems may generate outcomes that are unstable, inefficient, or even dangerous compared to individual systems, which appear to behave as intended in isolation.[62]

Importantly, these dynamics are not confined to highly speculative future scenarios.[63] Even current systems demonstrate early forms, or precursors, of these behaviors. Experts have highlighted how automated trading systems can interact to produce rapid market fluctuations.[64] Cyber tools can adapt to defenses in near real-time.[65] AI-assisted research pipelines can accelerate discovery processes while simultaneously introducing new forms of risk that are not fully understood. As these systems become more capable and more widely deployed, the likelihood of such interactions producing significant consequences increases.[66]

Emergence, opacity, and multi-agent interaction suggest that risk in AI-enabled environments may increasingly be driven by system behavior, as opposed to discrete inputs. This marks a significant departure from many traditional risk models: models that typically assume risks can be traced to identifiable sources and mitigated through targeted interventions. In contrast, these AI-enabled systems may produce risks that are distributed, dynamic, and difficult to attribute to any single component or actor.

Risk, in Theory

The preceding sections highlight that risk in AI-enabled environments is shaped by system behavior rather than discrete, siloed inputs. This raises a central question: if risk emerges from the interaction of complex, tightly coupled, and often opaque systems, how can we understand, manage, or mitigate such risks? Two frameworks, "normal accident theory" and "high-reliability organization theory", offer useful, if incomplete, lenses through which to examine this challenge.

Normal accident theory begins from a stark premise: accidents should be expected in systems whose components interact in complex, unexpected ways and whose processes are so tightly coupled that failures can cascade faster

60 Anthropic, "Agentic Misalignment: How LLMs Could be Insider Threats," Anthropic Research, June 20, 2025. https://www.anthropic.com/research/agentic-misalignment

61 For sources, please see Patricia Paskov, Jeffrey Lee, Kyle Brady, and Alyssa Worland, Measuring Biological Capabilities and Risks of AI Agents: Generating and Interpreting Evidence from Agentic Evaluations (Santa Monica, CA: RAND Corporation, 2026), https://www.rand.org/pubs/perspectives/PEA4710-1.html. Accessed March 30, 2026; Hao Li, Chunhong Guo, Peter Ping Li, and Fangbai Song, "AI Automation and AI Opacity: The Effects on Threat and Response Appraisal," Information and Management, Vol. 63, No. 3 (April 2026), https://doi.org/10.1016/j.im.2026.104314. Accessed March 30, 2026; and Chunhong Guo, Huifang Liu, Fangbai Song, and Jingfu Guo, "The Double-Edged Sword Effects of Algorithmic Opacity: The Self-Determination Theory Perspective," Acta Psychologica, Volume 260 (October 2025), https://doi.org/10.1016/j.actpsy.2025.105600. Accessed March 30, 2026.

62 Philipp Altmann et al., "Emergence in Multi-Agent Systems: A Safety Perspective," arXiv preprint arXiv:2408.04514, August 8, 2024: https://arxiv.org/html/2408.04514v1. Accessed March 30, 2026.

63 Matthew Hutson, "AI Agents Break Rules Under Everyday Pressure," IEEE Spectrum, January 21, 2026: https://spectrum.ieee.org/ai-agents-safety

64 Maximilian Goehmann, "AI and the Stock Market: Are Algorithmic Trades Creating New Risks?," Research for the World, London School of Economics and Political Science, September 23, 2025, https://www.lse.ac.uk/research/research-for-the-world/ai-and-tech/ai-and-stock-market. Accessed March 30, 2026.

65 Open AI, "Scaling Trusted Access for Cyber Defense," OpenAI, April 14, 2026, https://openai.com/index/scaling-trusted-access-for-cyber-defense/

66 National Academies of Sciences, Engineering, and Medicine. The Age of AI in the Life Sciences: Benefits and Biosecurity Considerations. (Washington, DC: The National Academies Press), 2025.

than operators can diagnose or contain them. In such environments, accidents are not aberrations, but features of the system itself. Multiple small failures can combine in ways that are difficult to predict, producing outcomes that do not necessarily stem from negligence or malicious intent. Rather, they emerge from structural properties of the system—properties that may become visible only under stress or at scale.[67]

This perspective has direct relevance for AI and AI-enabled systems (as it did for nuclear systems before them).[68] As discussed above, the integration of AI across domains introduces new forms of complexity, interdependence, opacity, and speed. Systems that were once loosely coupled become more tightly integrated through shared data pipelines, automated decision-support tools, and real-time feedback mechanisms. In such settings, failures may also no longer be localized, but may cascade across components, or organizational boundaries in ways that are difficult or impossible to anticipate or arrest.[69] Therefore, from a "normal accidents" perspective, AI could push systems into ones where certain classes of failure may become increasingly difficult to avoid.

High-reliability organization (HRO) theory, by contrast, offers a more optimistic view of how risk can be managed in safety-critical environments. Emerging from studies of organizations that operate hazardous systems such as nuclear power plants, aircraft carriers, and air traffic control systems, HRO theory suggests that it is possible to achieve consistently safe performance even under conditions that would otherwise present significant risks. However, this outcome depends on a set of organizational practices and cultural norms that prioritize vigilance, adaptability, and continuous learning from realized and potential failures.[70]

Applied to AI and AI-enabled systems, HRO Theory suggests that the risks introduced by complexity and coupling may be mitigated partially through deliberate organizational practices. This includes rigorous testing and evaluation, continuous monitoring of system performance, learning from mistakes, clear lines of accountability, the integration of human judgment in critical decision-making processes, and ensuring those closest to an activity or issue cease operations if problems emerge.[71] It also implies the need for institutions that are capable of *adapting* to evolving technologies, rather than relying on static rules or assumptions.

Both theories highlight key tensions that are relevant to this report. From a normal accident perspective, AI could introduce increased complexity and tight coupling in ways that may outpace the ability of organizations to fully understand or control the systems they operate. As these AI systems become more autonomous, more opaque, and more interconnected, this increases the likelihood of failures and raises the possibility that certain risks may be inherent to the system. In situations like this, improved design or governance is unlikely to eliminate such risks.

From an HRO perspective, the conditions required for high reliability, including organizational discipline, transparency, and a culture of safety, may be difficult to sustain in environments characterized by rapid innovation, competitive pressure, and fragmented governance. In particular, the diffusion of AI capabilities across a wide range of actors, including those outside traditional regulatory or institutional frameworks, complicates efforts to maintain consistent standards of reliability, let alone the norms of safety and security necessary for such cultures to be successful.

The normal accident and HRO perspectives highlight four key challenges for AI's impact on global risks. First, there is a challenge of scale. High-reliability practices have historically been developed within tightly controlled

67 Key foundational texts on this topic include the original work of Charles Perrow. This framework was then applied to Scott Sagan's Limits of Safety. Sources are Charles Perrow, Normal Accidents: Living with High-Risk Technologies (Basic Books: New York, 1984) and Scott D. Sagan, The Limits of Safety: Organizations, Accidents, and Nuclear Weapons (Princeton, NJ: Princeton University Press, 1995).

68 Sagan, The Limits of Safety (1995), pp. 1 - 30.

69 Federico Bianchi, Amanda Cercas Curry, and Dirk Hovy, "Viewpoint: Artificial Intelligence Accidents Waiting to Happen?" Journal of Artificial Intelligence, Vol. 76 (2023): pp. 193 - 199, https://jair.org/index.php/jair/article/view/14263

70 Todd R. LaPorte and Paula M. Consolini, "Working in Practice but Not in Theory: Theoretical Challenges of High-Reliability Organizations," Journal of Public Administration Research and Theory: J-PART, Vol. 1, No. 1 (January 1991): pp. 19 - 48. https://polisci.berkeley.edu/sites/default/files/people/u3825/LaPorte-WorkinginPracticebutNotinTheory.pdf. Accessed March 30, 2026.

71 Karl E. Weick and Kathleen M. Sutcliffe, Managing the Unexpected: Resilient Performance in an Age of Uncertainty, 3rd Edition, San Francisco: Jossey-Bass, 2013.

organizational environments. AI-enabled systems, by contrast, are often distributed across multiple organizations, jurisdictions, and domains. Further, as AI becomes more reliable, human oversight may actually be harder as humans in the loop grow complacent with AI outputs. This makes it more difficult to implement and enforce the practices required for reliability.[72]

Second, there is the challenge of speed. Both theories were developed in contexts where complex systems operated on timescales that allowed for some degree of human intervention and deliberation. AI-enabled systems, especially those operating in real-time or near real-time environments, compress decision timelines in ways that may reduce opportunities for detection, coordination, and response.

Third, there is the challenge of visibility. High-reliability organizations rely on the ability to observe and interpret system behavior, including weak signals of failure. The opacity of many AI systems complicates this process, making it more difficult to understand how decisions are being made or to identify emerging risks before they manifest.

Finally, there is the challenge of control. Both theories assume that human operators and institutions remain the ultimate authority over the systems they manage. As AI systems become more capable and more deeply embedded in critical functions, the locus of control may shift in ways that are not fully understood or easily reversed.

These views have helped shape existing frameworks such as NIST's risk management framework, which builds on the concept of iterative resilience: the use of cyclical and continuous processes to anticipate, withstand, and recover from adverse conditions in domains like AI and cyber.[73] Instead, they should be viewed as complementary perspectives that highlight both the inevitability of certain failures and the possibility of mitigating others through institutional design and practice.

From Systems to Actors

The analysis in this section suggests that global risk in the age of AI emerges from the interaction between complex systems, evolving capabilities, and institutional constraints. Traditional frameworks such as TVC remain useful for structuring analysis and identifying intervention points, but AI introduces new dynamics, including emergence, opacity, and multi-agent interaction, that complicate their underlying assumptions.

Existing theory offers important insights into how risk develops and how it may be mitigated in complex environments. But these theories also face limitations in systems characterized by rapid change, distributed control, and limited transparency. In such settings, risk cannot be understood solely by analyzing individual components in isolation. Rather, it must also be examined through system behavior and interaction.

At the same time, these dynamics ultimately manifest through actors: the people and institutions that design, deploy, govern, and attempt to exploit AI-enabled systems. Understanding global risk therefore requires closer attention to who these actors are, the incentives shaping their behavior, and how their actions influence the broader risk landscape.

72 For sources, please see "When Better AI Makes Oversight Harder," Wharton AI Analytics Initiative, 2025, https://ai-analytics.wharton.upenn.edu/insights/when-better-ai-makes-oversight-harder/ and Antti Salovaara, Kalle Lyytinen, and Esko Penttinen, "High Reliability in Digital Organizing: Mindlessness, the Frame Problem, and Digital Operations," MIS Quarterly, Vol. 43, No. 2 (2019): pp. 555 - 578. https://misq.umn.edu/misq/article/43/2/555/1646/High-Reliability-in-Digital-Organizing. Accessed March 30, 2026.

73 National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1 (Gaithersburg, MD: U.S. Department of Commerce, January 2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Chapter 3. Threat - Intent and Capability

The previous chapter introduced threat, vulnerability, and consequence as a way to decompose risk and identify points for intervention. This chapter examines threat in more detail: the actors, intentions, and capabilities that create pathways for harm.

Because threat analysis is already central to national security, this chapter draws on government and military examples to show where AI-related risks are emerging, how they may evolve, and where future capabilities could intensify existing dangers.

Threat can also be understood through a simple heuristic:

$$\textit{Threat} = \textit{Intent} \times \textit{Capability}$$

AI can shape both sides of this equation. It can increase capability by helping actors scale operations, process larger datasets, personalize targeting, and act at machine speed. It can also complicate intent, especially when decisions are delegated to systems whose behavior may be misinterpreted, inferred from outputs, or only partially understood.

For those in the autonomous power camp, the concern goes further. Sufficiently advanced systems could, themselves, become threat actors with emergent goals, strategic behavior, and capabilities beyond human control. The sections below examine how AI changes the threat landscape across actors, incentives, and systems.

Actors and Intent: Who Generates AI-Related Threats?

Understanding AI-related threats to international security requires looking across a wide range of actors: states, militaries, intelligence agencies, AI developers, criminal groups, terrorist organizations, and other actors that may misuse AI capabilities. Major military powers remain central because they have the resources, doctrine, and the institutional pathways to integrate AI into high-consequence systems. However, AI capabilities are also diffusing through commercial, academic, and open-source ecosystems.[74] As a result, threat analysis cannot assume a small number of rational, unitary state adversaries. It must account for a heterogeneous set of actors operating with different incentives, time horizons, and accountability structures.

At the state level, major powers are already driving the frontier of AI-enabled military integration (though with friction common to large, entrenched bureaucracies),[75] which includes embedding these technologies into intelligence, surveillance, command and control,[76] cyber operations, and targeting pipelines.[77] Their incentives are

74 Office of the Director of National Intelligence, Annual Threat Assessment of the U.S. Intelligence Community, 2026 (Washington, DC: ODNI, March 2026), https://www.dni.gov/files/ODNI/documents/assessments/ATA-2026-Unclassified-Report.pdf

75 Zach Hughes, "Fog, Friction, and Thinking Machines," War on the Rocks, March 11, 2020, https://warontherocks.com/2020/03/fog-friction-and-thinking-machines/ and Michael Raska and Richard A. Bitzinger, eds. The AI wave in defence innovation: Assessing military artificial intelligence strategies, capabilities, and trajectories. Taylor & Francis, 2023.

76 Schmidt, Eric, Robert Work, Safra Catz, Eric Horovitz, Steve Chien, Andrew Jassy, Mignon Clyburn et al. "National Security Commission on Artificial Intelligence (AI)," 2021.

77 Yuval Abraham, "Lavender: The AI Machine Directing Israel's Bombing Spree in Gaza," +972 Magazine, April 3, 2024, https://www.972mag.com/lavender-ai-israeli-army-gaza/; Heidy Khlaaf, Sarah Myers West, and Meredith Whittaker, "Mind the Gap: Foundation Models and the Covert Proliferation of Military Intelligence, Surveillance, and Targeting," arXiv preprint arXiv:2410.14831, October 18, 2024, https://arxiv.org/abs/2410.14831; and Parmy Olson, "Claude AI Helped Bomb Iran. But How Exactly?," Bloomberg Opinion, March 4, 2026, https://www.bloomberg.com/opinion/articles/2026-03-04/iran-strikes-anthropic-claude-ai-helped-us-attack-but-how-exactly

shaped by a revival of great power competition, concerns about strategic advantage (and an attendant "race to the bottom"), and fears of technological surprise.[78]

For U.S. policymakers, China looms large.[79] Washington has used export controls on advanced semiconductors and manufacturing equipment to limit China's access to frontier AI compute. Beijing, in turn, has accelerated efforts to build a more self-sufficient AI ecosystem around domestic chipmakers, cloud providers, and model developers.[80]

The results remain contested. Chinese firms still face compute constraints, and U.S. models retain important advantages across many frontier benchmarks.[81] At the same time, companies like Huawei and DeepSeek have become important symbols of China's effort to reduce dependence on U.S. technology and remain competitive in advanced AI development.[82] From the U.S. perspective, the competition centers on leadership in compute, models, talent, standards, and supply chains. From the Chinese perspective, AI is both an engine of economic modernization and a domain of technological sovereignty.

These dynamics create pressure for rapid adoption, especially when rivals appear to be moving quickly. Regional and middle powers face different constraints. Some see AI as a way to offset conventional disadvantages, while others use it to signal technological sophistication or strengthen deterrence.[83] Ukraine's use of drones and Russia's rapid development of autonomous capabilities show how battlefield pressures can accelerate experimental adoption.[84],[85]

Outside active conflicts, states are also deciding where to position themselves in the AI supply chain, from semiconductor manufacturing to sovereign model development.[86] In both cases, adoption may be opportunistic and uneven, increasing the risk that capabilities outpace the institutions needed to govern them responsibly.

Beyond states, the expansion of AI capabilities has lowered barriers to entry for a wider range of non-state actors. Criminal syndicates may use AI to scale cyber exploitation, automate fraud, or enhance illicit logistics.[87] Terrorist organizations and other ideologically motivated groups may use AI for propaganda, recruitment, targeting, or

78 R. Evan Ellis, "Race to the Bottom: China and the Self-Defeating Logic of Transactional Diplomacy in the Americas," The Diplomat, April 18, 2023, https://thediplomat.com/2023/04/race-to-the-bottom-china-and-the-self-defeating-logic-of-transactional-diplomacy-in-the-americas/

79 William Hannas and Huey-Meei Chang, "China's Artificial General Intelligence: We're Still Getting it Wrong," Center for Security and Emerging Technology, August 29, 2025, https://cset.georgetown.edu/article/chinas-artificial-general-intelligence/; Kyle Chan, Gregory Smith, Jimmy Goodrich, Gerard DiPippo, and Konstantin F. Piz, "Full Stack: China's Evolving Industrial Policy for AI," RAND, June 26, 2025, https://www.rand.org/pubs/perspectives/PEA4012-1.html; and Cole McFaul, Sam Bresnick, and Daniel Chou, "Pulling Back the Curtain on China's Military-Civil Fusion," Center for Security and Emerging Technology, September 2025, https://cset.georgetown.edu/article/how-china-is-using-ai-to-win-future-wars/;

80 Nathan Lambert, "Notes from Inside China's AI Labs," Interconnects (Substack), April 2026, https://www.interconnects.ai/p/notes-from-inside-chinas-ai-labs?utm_source=substack&utm_medium=email

81 Ritwik Gupta, Leah Walker, and Andrew W. Reddie, "Whack-a-Chip: The Futility of Hardware-Centric Export Controls," arXiv preprint arXiv:2411.14425, November 21, 2024, https://arxiv.org/abs/2411.14425

82 Chris McGuire, Michael C. Horowitz, and Jessica Brandt, "DeepSeek V4 Signals a New Phase in the U.S.-China AI Rivalry," Council on Foreign Relations, April 29, 2026, https://www.cfr.org/articles/deepseek-v4-signals-a-new-phase-in-the-u-s-china-ai-rivalry

83 Interestingly, there is significant variation among those countries focused on the front end (e.g., model development) and the back end (e.g., semiconductor supply chain, GPUs) of the AI supply chain.

84 Tereza Pultarova, "Autonomous Drone Warfare is Already Here," IEEE Spectrum, December 9, 2025, https://spectrum.ieee.org/autonomous-drone-warfare

85 Kateryna Bondar, "Russia is Building Tomorrow's War Machine," New York Times, April 21, 2026, https://www.nytimes.com/2026/04/21/opinion/russia-drones-putin-ukraine-war.html?unlocked_article_code=1.clA.emYW.uZBEHyx-UFcs&smid=nytcore-android-share

86 G42, "Abu Dhabi Launches Comprehensive Global Investment Strategy for Artificial Intelligence," G42 News, September 2025, https://www.g42.ai/resources/news/abu-dhabi-launches-comprehensive-global-investment-strategy-artificial-intelligence

87 For sources, please see Anthropic, "Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign," Anthropic News, November 13, 2025, https://www.anthropic.com/news/disrupting-AI-espionage and Anthropic, "Detecting and Countering Misuse of AI: August 2025," Anthropic News, August 27, 2025, https://www.anthropic.com/news/detecting-countering-misuse-aug-2025

surveillance.[88] Even relatively unsophisticated actors can now access tools for analysis, content generation, and operational planning that were once largely restricted to well-resourced institutions. This does not eliminate the importance of state power, but it does create a more crowded and unpredictable threat landscape in which capabilities are less tightly coupled to traditional indicators of strength.

Private firms and open-source research communities now occupy an increasingly central role in developing and disseminating AI technologies. Many of the most advanced systems are produced outside government, often with dual-use applications and global user bases.

This creates a tension where innovation is driven by decentralized commercial and research ecosystems, while risk management and national-security oversight remain primarily state responsibilities.[89] This also means incentives differ across actors. Firms may prioritize performance, market share, or deployment speed, while open-source communities often emphasize accessibility and experimentation. This dynamic complicates efforts to control diffusion, enforce standards, or attribute responsibility when systems are misused.

These trends also create hybrid actor configurations long familiar in cyber operations: state-backed proxies, private contractors, volunteer networks, and loosely affiliated technical communities.[90] States, therefore, may deliberately obscure their involvement by relying on commercially available tools or outsourcing activities to intermediaries, enabling plausible deniability. Conversely, non-state actors may benefit indirectly from capabilities that diffuse through global markets or are repurposed from state-developed systems. The result is a blurring of boundaries between state and non-state action that further complicates attribution, deterrence, and response.

AI systems can shape outcomes in ways that mediate, distort, or amplify the intent of human actors. For example, decision-support systems may prioritize certain courses of action based on training data or optimization criteria that are not fully transparent to operators.[91] Generative models may synthesize intelligence assessments that carry an aura of objectivity, influencing decision-makers even when underlying assumptions are flawed. In this context, AI does not replace human agency, but it reconfigures how intent is formed, interpreted, and enacted. Human actors still choose, but they increasingly do so through systems that shape what appears salient, feasible, urgent, or justified.[92]

Crucially, the most significant threats arise not from any single actor operating in isolation, but from the interaction effects among them. Capabilities diffuse across actors with different incentives. Actions taken by one actor may be misinterpreted by another. Feedback loops can also emerge as systems respond to one another's outputs. Similarly, the proliferation of AI-enabled cyber tools may generate a more contested and ambiguous environment in which attribution is difficult, and escalation thresholds are unclear.

88 C. Anthony Pfaff, The Weaponization of AI: The Next Stage of Terrorism and Warfare (Ankara: Centre of Excellence Defence Against Terrorism, 2025), https://www.tmmm.tsk.tr/publication/researches/21-TheWeaponizationofAI-TheNextStageofTerrorismandWarfare.pdf and United Nations Interregional Crime and Justice Research Institute and United Nations Counter-Terrorism Centre, Algorithms and Terrorism: The Malicious Use of Artificial Intelligence for Terrorist Purposes (Turin: UNICRI and UNCCT, 2021), https://unicri.org/News/Algorithms-Terrorism-Malicious-Use-Artificial-Intelligence-Terrorist-Purposes

89 U.S. Department of Commerce's Bureau of Industry and Security (BIS), "Framework for Artificial Intelligence Diffusion," Federal Register 90, No. 10 (January 15, 2025): 4543 - 4568, https://www.federalregister.gov/documents/2025/01/15/2025-00636/framework-for-artificial-intelligence-diffusion

90 Jamie Collier. "Proxy Actors in the Cyber Domain: Implications for State Strategy." St Antony's International Review 13, No. 1 (2017): 25-47; Florian J. Egloff, Semi-State Actors in Cybersecurity. Oxford University Press, 2021; Myriam Dunn Cavelty, "Cyber-Security and Private Actors." In Routledge Handbook of Private Security Studies, pp. 89-99. Routledge, 2015; and Tim Maurer. "'Proxies' and Cyberspace." Journal of Conflict and Security Law 21, No. 3 (2016): 383-403.

91 Yasir Atalan, Ian Reynolds, and Benjamin Jensen, "AI Biases in Critical Foreign Policy Decisions," Center for Strategic and International Studies, February 26, 2025, https://www.csis.org/analysis/ai-biases-critical-foreign-policy-decisions

92 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. 2 (2024): https://academic.oup.com/isq/article/68/2/sqae020/7638566

AI as a Threat Actor

Some analysts argue that this human-centered framing of intent and capability may become less applicable as AI systems gain greater autonomy and strategic flexibility. This concern is strongest in the "autonomous power" camp, which treats advanced AI not merely as a tool that shapes human intent and capability, but as a possible actor whose own objectives become increasingly strategically relevant.[93] On this view, the key question is not only what humans want to do with AI, but what increasingly autonomous systems do when given broad goals, long time horizons, and freedom to choose how to pursue them. While this possibility is most salient in debates over artificial general intelligence and superintelligence, even today's early AI agents offer a limited preview of what it can look like for humans to specify a high-level objective while leaving the system to determine the intermediate steps.[94]

One way researchers explain how AI could become a threat actor is through the idea of *convergent instrumental goals*: the possibility that highly capable AI systems pursuing very different objectives from their human controllers may, nevertheless, adopt similar intermediate strategies because those strategies improve the likelihood of success across many goals.[95] The potential threat then emerges from AI's pursuing harmful sub-goals in a way that escapes human control.

Researchers have presented controlled demonstrations that offer early, limited evidence of this dynamic.[96] In some settings, models given broad goals and opportunities to act against human instructions have selected harmful intermediate actions, including manipulating evaluation conditions or resisting shutdown.

However, interpretation of these results remains contested. Many demonstrations rely on artificial environments, explicit goal prompts, or evaluation setups that may incentivize scheming-like behavior. Critics also argue that the field should distinguish more carefully between showing that a model can produce such behavior under specific conditions and showing that it is likely to do so spontaneously in real-world deployments.[97]

Concerns about AI as a potential threat actor become more significant as systems grow more capable, autonomous, and general-purpose. Greater capability does not automatically produce harmful intent, but it can allow systems to pursue objectives more effectively over longer time horizons and across more complex environments. Therefore, some researchers argue that sufficiently capable and misaligned systems could eventually pose risks on an existential scale.[98]

93 Stuart Russell, Human Compatible: Artificial Intelligence and the Problem of Control. New York: Viking, 2019, https://www.penguinrandomhouse.com/books/566677/human-compatible-by-stuart-russell/ and Yoshua Bengio, "How Rogue AIs May Arise," May 22, 2023, https://yoshuabengio.org/en/blog/how-rogue-ais-may-arise

94 Yoshua Bengio et al., International AI Safety Report 2026, DSIT 2026/001, February 3, 2026, https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026.

95 Stephen M. Omohundro, "The Basic AI Drives," in Artificial General Intelligence 2008: Proceedings of the First AGI Conference, ed. Pei Wang, Ben Goertzel, and Stan Franklin, Frontiers in Artificial Intelligence and Applications 171 (Amsterdam: IOS Press, 2008), 483 - 492, https://selfawaresystems.com/wp-content/uploads/2008/01/ai_drives_final.pdf; Nick Bostrum, "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents," Minds and Machines 22, No. 2 (2012): 71 - 85, https://nickbostrom.com/superintelligentwill.pdf; and Tsvi Benson-Tilsen and Nate Soares, "Formalizing Convergent Instrumental Goals," in Workshops at the Thirtieth AAAI Conference on Artificial Intelligence (Palo Alto: Association for the Advancement of Artificial Intelligence, 2016), https://intelligence.org/files/FormalizingConvergentGoals.pdf

96 Anthropic, "Agentic Misalignment: How LLMs Could Be Insider Threats," Anthropic Research, June 20, 2025, https://www.anthropic.com/research/agentic-misalignment;

97 Arvind Narayanan and Sayash Kapoor, "AI as Normal Technology," AI as Normal Technology (Substack), April 15, 2025, https://www.normaltech.ai/p/ai-as-normal-technology and Christopher Summerfield et al., "Lessons from a Chimp: AI 'Scheming' and the Quest for Ape Language," UK AI Security Institute, July 4, 2025, https://www.aisi.gov.uk/research/lessons-from-a-chimp-ai-scheming-and-the-quest-for-ape-language

98 Joseph Carlsmith, "Is Power-Seeking AI an Existential Risk?" arXiv preprint arXiv: 2206.13353, 2022, https://arxiv.org/abs/2206.13353; Center for AI Safety, "Statement on AI Risk," May 30, 2023, https://aistatement.com/; and Dan Hendrycks, Mantas Mazeika, and Thomas Woodside, "An Overview of Catastrophic AI Risks," arXiv, last revised October 9, 2023, https://doi.org/10.48550/arXiv.2306.12001.

The Drivers of Threat: Heterogeneous Capabilities

Capabilities also vary significantly across threat actors. AI-enabled threats do not manifest uniformly across actors or domains. Instead, they reflect a distribution of capabilities shaped by access to data, compute, talent, and operational integration. States, firms, and non-state actors differ in how they deploy AI across the kill chain. Across all categories, however, AI primarily functions as a capability multiplier, enhancing existing tools of coercion rather than introducing wholly novel forms of violence. Below, we examine this variation in capabilities across a wide variety of threat actors and across domains that reflect discussions across a number of our FAS workshops.[99]

AI-ENABLED CYBER, INFORMATION, AND COERCIVE THREATS IN THE GRAY ZONE

Across cyber, information, and gray zone contexts, AI enables more persistent, scalable, and precise forms of operations.[100] Large-scale data ingestion and pattern recognition improve reconnaissance and target identification. Automated exploit generation, tailored phishing, and scalable content production reduce the cost of attack. And adaptive systems can also respond to defense in near real time, refining future operations through continuous feedback.

In practice, this produces several interrelated effects. Automated cyber exploitation increases the scale and persistence of attacks by allowing systems to identify and exploit vulnerabilities across vast networks. Put simply, the floor for conducting sophisticated cyber operations has fallen. Additionally, generative systems also transform disinformation and influence campaigns by enabling adaptive, audience-specific content production at scale.[101] At the same time, integrating diverse data sources allows actors to target individuals, infrastructure, and institutions with increasing precision.

The military significance of these capabilities lies not in their novelty but in their speed, scale, and personalization. Actors can sustain continuous campaigns of disruption and influence, often below the threshold of armed conflict. These dynamics strain traditional deterrence and response frameworks, which rely on clear attribution, identifiable thresholds, and discrete acts of aggression.

This same logic extends, albeit with different technical constraints, to biological and radiological pathways. In biological contexts, AI can accelerate elements of the kill chain, including target identification and aspects of agent design, lowering informational barriers even as experimental constraints remain significant. In radiological and nuclear-adjacent contexts, AI may not currently enable weapon construction per se, but can support targeting, delivery optimization, and consequence modeling, thereby enhancing the effectiveness of coercive threats involving existing materials or systems. In both cases, the principal effect is not immediate transformation, but the expansion and integration of threat capabilities across domains.

99 For resources, please see Federation of American Scientists, Artificial Intelligence and Military Integration: Emerging Risks, Governance Challenges, and Strategic Stability (Washington, DC: Federation of American Scientists, December 2025) https://fas.org/wp-content/uploads/2025/12/1215-ai-mil.pdf and Federation of American Scientists, Artificial Intelligence, Cyber, and Global Risk: Current Status and Future Risks (Washington, DC: Federation of American Scientists, January 2026), https://fas.org/wp-content/uploads/2026/01/January-2026-AI-Cyber-Global-Risk.pdf

100 James Johnson. "The AI-Cyber Nexus: Implications for Military Escalation, Deterrence and Strategic Stability." Journal of Cyber Policy 4, No. 3 (2019): 442-460.

101 Noémi Bontridder and Yves Poullet. "The Role of Artificial Intelligence in Disinformation." Data & Policy 3 (2021); Todd C. Helmus, "Artificial Intelligence, Deepfakes, and Disinformation: A Primer." (2022); Tiffany Hsu and Stuart A. Thompson, "Disinformation Researchers Raise Alarms About A.I. Chatbots," New York Times, February 8, 2023, https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html; and Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova, "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations," arXiv preprint arXiv:2301.04246v1, 2023, https://arxiv.org/pdf/2301.04246.pdf

SIGNALING, SECRECY, AND MISPERCEPTION

A distinctive feature of AI-enabled capabilities is their opacity. Unlike traditional military assets, AI systems are difficult to observe, evaluate, and verify from the outside. This creates incentives for both states and firms to shape perceptions of their capabilities. Some actors may overstate levels of automation or system performance to enhance deterrence, signal technological sophistication, or attract investment, while others may obscure real capabilities to preserve operational advantage.

This dynamic produces a fundamental problem of unverifiable assurances. Claims of human-in-the-loop control (one of the most commonly proposed confidence-building measures[102]), for example, are difficult to substantiate in practice, particularly in domains such as intelligence, surveillance, and reconnaissance, command and control, and targeting. External observers—and in some cases even internal decision-makers—may lack visibility into how AI systems are actually used in operational settings.

The result is a risk environment in which beliefs about AI capabilities may matter more than their actual performance. In crisis contexts, worst-case assumptions about adversary capabilities may drive escalation, even if those capabilities are overstated, unreliable, or poorly integrated. AI thus amplifies a familiar problem in international security—misperception[103]—but does so in ways that are harder to correct through traditional signaling or verification mechanisms.

MANIPULATING AI SYSTEMS

More capable actors, particularly states and well-resourced non-state groups, are not limited to using AI as a tool. They can also target and manipulate the AI systems used by others,[104] introducing a new layer of competition focused on the integrity of data and models (also described as "adversarial applications").

One vector is training data poisoning,[105] which involves the subtle manipulation of datasets used to train models for intelligence, surveillance, and reconnaissance, targeting, logistics, or early warning. Such interventions may create latent effects that remain undetected until crisis or conflict conditions trigger system failures.

A second vector is operational data manipulation. By feeding adversary AI systems misleading or biased real-time inputs, such as corrupted sensor feeds or manipulated open-source intelligence streams, actors can degrade the performance of systems that rely on the fusion of diverse data sources. These effects can cascade through decision-support architectures, producing flawed assessments and recommendations.

A related strategy is information flooding. By overwhelming AI-enabled analytic systems with low-quality, misleading, or strategically crafted data, adversaries can distort how systems prioritize information and allocate attention. In systems that rely on pattern recognition and ranking, such manipulation can redirect attention away from genuinely salient signals and toward adversary-selected noise.

These risks are particularly acute in state-to-state contexts, where AI systems may support crisis management and decision-making. Even when humans remain formally in the loop, AI-generated outputs can exert disproportionate

102 Allan Dafoe. "AI Governance: a Research Agenda." Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK 1442 (2018): 1443; Sarah Shoker, Andrew Reddie, Sarah Barrington, Ruby Booth, Miles Brundage, Husanjot Chahal, Michael Depp et al. "Confidence-Building Measures for Artificial Intelligence: Workshop Proceedings." arXiv preprint arXiv:2308.00862 (2023); Michael C. Horowitz and Lauren Kahn. "Leading in Artificial Intelligence through Confidence Building Measures." The Washington Quarterly 44, no. 4 (2021): 91-106.

103 Robert Jervis. Perception and Misperception in International Politics: New Edition. Princeton University Press, 2017; and Janice Gross Stein. "Building Politics into Psychology: The Misperception of Threat." Political Psychology (1988): 245-271.

104 Shlomit Wagman and Sarah Hubbard, "Weaponized AI: A New Era of Threats and How We Can Counter It," Ash Center for Democratic Governance and Innovation, April 8, 2025, https://ash.harvard.edu/articles/weaponized-ai-a-new-era-of-threats/

105 Christopher Whyte, "Problems of Poison: New Paradigms and 'Agreed' Competition in the Era of AI-Enabled Cyber Operations." In 2020 12th International Conference on Cyber Conflict (CyCon), Vol. 1300, pp. 215-232. IEEE, 2020.

influence due to their perceived objectivity, speed, and perceived technical authority.[106] As a result, manipulation of AI systems becomes a pathway for shaping adversary perceptions and, potentially, their strategic choices.

Even the suspicion that an AI-enabled intelligence system has been compromised can create a kind of "Liar's dividend" in which states fear adversaries may possess capabilities such as data poisoning even if such capabilities are not used.

In operational settings, competition increasingly takes the form of autonomy–counter–autonomy dynamics, as actors seek both to deploy AI-enabled systems and to degrade or exploit those of their adversaries. Actors may directly exploit autonomous and semi-autonomous systems by spoofing visual, infrared, or radar signatures to confuse computer vision models, or by exploiting assumptions embedded in navigation, target recognition, or sensor fusion algorithms. Electronic warfare and GPS denial further complicate this environment by forcing adversaries to rely more heavily on autonomy at the edge. As systems degrade under contested conditions, operators may defer more to automated functions, introducing new and poorly understood failure modes. Adversaries can then adapt to exploit these behaviors, creating a feedback loop of action and counteraction.

These dynamics are already observable in contemporary conflicts, including Ukraine,[107] and are likely to generalize to future peer conflicts. They drive rapid deployment of AI-enabled systems under operational pressure, often without corresponding advances in testing, evaluation, and governance.

CAPABILITY DIFFUSION AND PERSISTENT COMPETITION

Together, these patterns suggest that AI is reshaping not only the capabilities of individual actors but the structure of competition itself. Capabilities once concentrated among advanced states are increasingly diffused, while more sophisticated actors develop methods to manipulate and degrade the systems of others.

The result is an environment characterized by persistent, low-level competition across domains, ambiguity in both capability and intent, and increasing reliance on AI-mediated perception and decision-making.

Reframing "Threat" for the AI Era

Reflecting on both intent and capability, the growing risk associated with AI–military integration potentially lies less in autonomy as a discrete technological feature and more in the gradual erosion of control, understanding, and associated restraint under conditions of strategic competition. As actors face pressure to adopt and deploy AI-enabled capabilities, they may do so in ways that outpace their ability to comprehend system behavior, maintain meaningful oversight, or adhere to established norms of caution—made more acute given the nascent rise of agentic systems. The resulting dynamics are not driven solely by what these systems can do, but by how they are integrated into organizational processes and decision-making under time pressure and uncertainty.

Accordingly, assessing the threat posed by AI requires shifting analytical focus away from the technology in isolation and toward the broader systems in which it is embedded. This includes the incentives that drive adoption and use, the institutional contexts that shape implementation and oversight, the nature of human–machine interaction in operational settings, and the pathways through which these factors may contribute to escalation.[108] Together, these elements determine not only the capabilities available to actors but also how those capabilities are interpreted, trusted, and acted upon in practice.

106 The Future of Life Institute's "Artificial Escalation" does a particularly good job of demonstrating these concerns in a short video: https://futureoflife.org/project/artificial-escalation/

107 Michael C. Horowitz, Shira Pindyck, and Lauren Kahn. "The Drone Debates: Rethinking How Emerging Military Technologies Shape Power." Available at SSRN (2026). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6308358

108 Michael Mayer. "Trusting Machine Intelligence: Artificial Intelligence and Human-Autonomy Teaming in Military Operations." Defense & Security Analysis 39, No. 4 (2023): 521-538.

Threats do not become risks in isolation. They become consequential when they encounter systems that are brittle and unable to absorb failure. The next chapter therefore turns from who or what may generate AI-related threats to where those threats can take hold: the technical, organizational, and societal vulnerabilities that shape whether AI-enabled capabilities and malicious intent remain manageable or become sources of systemic harm.

Chapter 4. Vulnerability - An Architecture of Fragility

Having outlined the intent and capabilities of relevant threat actors, we now turn to the vulnerabilities that determine whether those threats can materialize. As in the previous chapter, we continue to focus on AI's use in military contexts because these environments concentrate many of the features that make AI-related risks especially consequential: high-stakes decisions, compressed timelines, adversarial pressure, complex human-machine systems, and limited tolerance for failure.

This chapter examines vulnerability not as an incidental feature of AI-enabled military systems, but as a structural property of the architectures through which these systems are developed and deployed.[109] If the preceding chapter focused on threat, or what actors do, this chapter turns to the systems through which those threats manifest and propagate. This shift highlights a critical distinction: threats may be external, but vulnerability is endogenous. It reflects choices made by system designers, operators, and institutions. In that sense, vulnerability is often the dimension defenders can shape most directly.

To understand vulnerability in this context, it is useful to think in terms of layered systems. Complex military architectures, particularly those integrating AI, resemble the "Swiss Cheese" model commonly used in safety engineering: a framework in which multiple layers of defense each contain latent weaknesses.[110] Failures occur not because of a single layer collapse, but because weaknesses align across layers, allowing hazards to propagate through the system.[111]

Crucially, vulnerability in AI-enabled systems does not arise solely from AI model error in the narrow sense. It emerges when control is attenuated through delegation and tight coupling, when visibility is degraded by opacity and incomplete testing, and when institutional restraint weakens under pressures of speed, competition, and technological diffusion.[112] These dynamics are not incidental. They are structural and systemic, reflecting how AI is integrated into broader sociotechnical systems. As such, vulnerability is not reducible to isolated technical flaws, but is a property of the system as a whole.

Control, Visibility, and Restraint

At the core of this architecture of fragility are three interacting dynamics referenced in the prior chapters: the attenuation of control, the degradation of visibility, and the erosion of institutional restraint.

Control is attenuated as tasks are delegated to increasingly complex and tightly coupled systems. Drawing on insights from complexity theory and normal accident theory,[113] tightly coupled systems reduce the time and space available for human intervention.[114]

109 Sven Fuchs, Jörn Birkmann, and Thomas Glade. "Vulnerability Assessment in Natural Hazard and Risk Analysis: Current Approaches and Future Challenges." Natural Hazards 64, No. 3 (2012): 1969-1975.

110 Gwendolyn CH Bakx and James M. Nyce. "Risk and Safety in Large-Scale Socio-Technological (military) Systems: a Literature Review." Journal of Risk Research 20, No. 4 (2017): 463-481.

111 Gianluca Pescaroli, Robert T. Wicks, Giampierro Giacomello, and David E. Alexander. "Increasing Resilience to Cascading Events: The M. OR. D. OR. Scenario." Safety Science 110 (2018): 131-140.

112 Toni Erskine 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; Benjamin M. Jensen, 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.

113 Charles Perrow. "Normal Accidents: Living with High Risk Technologies-Updated Edition." (2011): 1-464.

114 Jianguo Liu, Thomas Dietz, Stephen R. Carpenter, Marina Alberti, Carl Folke, Emilio Moran, Alice N. Pell et al. "Complexity of Coupled Human and Natural Systems." Science 317, No. 5844 (2007): 1513-1516.

When AI systems are embedded in decision-support or operational pipelines, actions may be triggered or recommended faster than human operators can fully interrogate them.[115] Delegation is about assigning tasks to machines *and* reconfigures authority and agency within the system itself.[116] As delegation increases, the locus of control becomes more diffuse, which impairs the human's ability to intervene meaningfully.

Visibility also degrades as systems become more opaque and difficult to interpret. Many AI models, particularly those used for perception, classification, or prediction, still cannot provide transparent accounts of how outputs are generated. This complicates both real-time decision-making and accountability.[117] Operators may over- or under-trust system outputs, neither of which supports effective oversight. Calibrated trust is, itself, an evaluation problem: operators need reliable signals about when systems are likely correct. It also undermines confidence calibration: operators may either over-trust or under-trust system outputs, neither of which is conducive to effective oversight. These issues are further exacerbated by limited testing. Systems are often evaluated under controlled settings that fail to capture the adversarial, dynamic, and uncertain environment in which they ultimately operate.

Institutional restraint, meanwhile, might weaken under the pressures of competition and speed. As actors "race to the bottom" to adopt and deploy AI capabilities,[118] organizational incentives may prioritize rapid integration over careful validation. Norms of caution—particularly in high-stakes domains—may erode as the perceived costs of falling behind increase. Diffusion further complicates this dynamic, as capabilities spread across actors with varying levels of institutional maturity and governance capacity.

Together, these dynamics create conditions in which systems are more difficult to control, harder to understand, and more likely to be used in ways that exceed intended bounds. We draw on examples from military and government contexts in this chapter to illustrate how these pathologies drive vulnerability.

Technical Vulnerabilities

At a technical level, AI-enabled systems introduce vulnerabilities that expand attack surfaces and make system integrity harder to verify. Techniques such as sensor spoofing, data poisoning, and prompt injection allow adversaries to manipulate inputs in ways that produce misleading or harmful outputs.[119] These attacks exploit AI systems' dependence on data rather than targeting physical infrastructure directly, with model opacity further compounding these vulnerabilities.

115 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. 2 (2024): sqae020; and see also, https://futureoflife.org/project/artificial-escalation/

116 J. Christopher Brill, M. L. Cummings, A. W. Evans III, Peter A. Hancock, Joseph B. Lyons, and Kevin Oden. "Navigating the Advent of Human-Machine Teaming." In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 62, No. 1, pp. 455-459. Sage CA: Los Angeles, CA: SAGE Publications, 2018; and Michael Mayer. "Trusting Machine Intelligence: Artificial Intelligence and Human-Autonomy Teaming in Military Operations." Defense & Security Analysis 39, No. 4 (2023): 521-538.

117 Jie Guo. "The Ethical Legitimacy of Autonomous Weapons Systems: Reconfiguring War Accountability in the Age of Artificial Intelligence." Ethics & Global Politics 18, No. 3 (2025): 27-39; Sebastian Bordt, Michèle Finck, Eric Raidl, and Ulrike Von Luxburg. "Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts." In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 891-905.

118 Colin H. Kahl, "The Myth of the AI Race: Neither America Nor China Can Achieve True Tech Dominance," Foreign Affairs, January 12, 2026, https://www.foreignaffairs.com/united-states/myth-ai-race; Denise Garcia, The AI Military Race: Common Good Governance in the Age of Artificial Intelligence. Oxford University Press, 2024; Jeffrey Ding, Technology and the Rise of Great Powers: How Diffusion Shapes Economic Competition; Princeton University Press, 2024; and David Brooks, "Trump is Winning the Race to the Bottom," The New York Times, July 17, 2025, https://www.nytimes.com/2025/07/17/opinion/trump-america-china.html

119 Christopher Whyte, "Problems of Poison: New Paradigms and 'Agreed' Competition in the Era of AI-Enabled Cyber Operations." In 2020 12th International Conference on Cyber Conflict (CyCon), Vol. 1300, pp. 215-232. IEEE, 2020 and OpenAI, OpenAI o1 System Card (San Francisco: Open AI, December 5, 2024), https://cdn.openai.com/o1-system-card-20241205.pdf

The limitations of benchmarking are particularly salient in this context. Benchmarking regimes typically measure performance under static and controlled conditions, providing limited insight into how systems will behave under adversarial pressure or in crisis scenarios. They rarely simulate deception, distributional shift, or strategic manipulation. As a result, systems that perform well in testing environments may exhibit brittle or unpredictable behavior in operational contexts. These limitations underscore the importance of evaluation regimes that account for adversarial behavior, uncertainty, and operational stress rather than relying solely on static benchmark performance.[120]

DATA INFRASTRUCTURE VULNERABILITIES

Beyond individual models, vulnerability is deeply embedded in the data infrastructures that underpin AI systems. Military applications, for example, rely on a heterogeneous mix of data sources, including commercial satellite feeds, open-source data streams, contractor-managed datasets, and contributions from allied partners. While this diversity enhances capability, it also introduces multiple points of fragility. If things go wrong, it can be extremely difficult to identify which data sources contributed to the error.

Contamination at ingestion points, in particular, represents a key vulnerability. Data may be corrupted, manipulated, or selectively curated before entering the system, introducing biases or distortions that propagate through downstream processes.[121] These issues are often difficult to detect, particularly when data sources are external or proprietary.

Hidden biases introduced upstream further complicate matters. For example, training data can include systematic distortions that shape model outputs in subtle but consequential ways. These biases may remain latent until specific operational conditions trigger their effects.

Dependency on non-sovereign data providers introduces an additional layer of risk. Reliance on commercial or allied sources creates exposure to supply disruptions, political constraints, and potential manipulation. Even in the absence of adversarial interference, operational environments evolve over time, producing what might be termed "out-of-sample risk." Systems trained on historical data may perform poorly when confronted with novel conditions, particularly in dynamic conflict environments.

THIN EVIDENCE BASE

A related and often underappreciated vulnerability stems from the thin empirical foundation underlying many high-stakes AI deployments. Unlike other domains of innovation—where decades of operational experience, testing regimes, and doctrinal evolution provide a basis for validation—AI-enabled systems are frequently introduced into decision-making environments with limited historical precedent and weak empirical grounding. This is due to a mismatch between available data and the conditions under which systems are expected to perform.[122]

This problem is particularly acute at higher levels of the escalation ladder. There are few, if any, empirical cases involving AI-supported decision-making in nuclear or near-nuclear contexts that would allow for robust validation of system performance. Even in lower-intensity conflicts, the integration of AI into operational pipelines is recent and uneven, producing a fragmented and context-specific evidence base. As a result, models are often trained on data that is only weakly representative of the environments in which they will be used—whether due to differences

120 This insight is reflected in European efforts to distinguish, with difficulty, "high-risk" use cases. These can specifically be viewed at the European Union, "Annex III: High-Risk AI Systems Referred to in Article 6(2)," in Artificial Intelligence Act, accessed May 10, 2026, https://www.euaiact.com/annex/3

121 Mathew J. Walter, Aaron Barrett, and Kimberly Tam. "Preventing Adversarial AI Attacks against Autonomous Situational Awareness: A Maritime Case Study." arXiv preprint arXiv:2505.21609 (2025) and Alex Wilner and Casey Babb, "New Technologies and Deterrence: Artificial Intelligence and Adversarial Behaviour." In NL ARMS Netherlands Annual Review of Military Studies 2020: Deterrence in the 21st Century—Insights from Theory and Practice, pp. 401-417. The Hague: TMC Asser Press, 2020.

122 Yoshua Bengio et al., International AI Safety Report 2026, DSIT 2026/001, February 3, 2026, https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026.

in adversary behavior, operational tempo, or the strategic stakes involved. This makes out-of-sample use especially risky in high-stakes settings.

This mismatch introduces a form of epistemic fragility. Systems may perform well under training or test conditions yet fail in precisely those scenarios that matter most—novel, adversarial, and high-pressure environments. The absence of large-N empirical validation further constrains confidence. In many cases, validation relies on simulation, red-teaming, or extrapolation from adjacent domains, each of which carries its own limitations.[123] Simulations may fail to capture adversarial adaptation; red-teaming exercises are bounded by the imagination and incentives of participants; and extrapolation assumes representativeness of the scenarios that have been modeled (where they are rarely collectively exhaustive).

The thin evidence base also has implications beyond technical performance. It shapes how organizations develop doctrine and how decision-makers interpret system outputs. In the absence of well-established use cases, operators often lack clear mental models for when to trust or override AI-generated recommendations.[124] This ambiguity can produce both over-reliance and under-utilization, depending on context and institutional culture.

Perhaps more subtly, the lack of empirical grounding already encourages a reliance on analogy and narrative in place of evidence. Decision-makers may draw on historical parallels—whether to Cold War early warning systems or precision-guided munitions—that only partially map onto AI-enabled systems. These analogies can be useful heuristics, but they also risk obscuring key differences, particularly around system behavior under uncertainty and adversarial manipulation.

In this sense, the thin evidence base is not merely a gap to be filled over time. It is a structural feature of AI-enabled integration (particularly in military contexts where a significant amount of risk lies), reflecting the pace of technological change relative to the accumulation of operational experience. As such, it introduces uncertainty into both system design and strategic decision-making, reinforcing the broader architecture of fragility described in this chapter.

COMPUTE, INFRASTRUCTURE, AND DEPENDENCY RISKS

AI-enabled military systems are also vulnerable through their dependence on underlying compute, infrastructure, and supply chains. Unlike many legacy military capabilities, which are developed and sustained within relatively closed and sovereign systems, contemporary AI architectures are deeply embedded in commercial, globalized, and interdependent nodes. This dependence introduces new pathways through which disruption, degradation, or manipulation can occur.

Both governments and militaries increasingly rely on cloud service providers for storage, processing, and model deployment. These arrangements offer significant advantages in terms of scalability, flexibility, and access to cutting-edge capabilities. At the same time, they create exposure to risks that are only partially under their control. As observed in both Ukraine and Iran, physical infrastructure—such as data centers, fiber-optic networks, and undersea cables—becomes a potential target for sabotage or disruption.[125] Cyber vulnerabilities within cloud environments introduce additional avenues for compromise, particularly in multi-tenant architectures where isolation is imperfect. Supply chain dependencies further complicate matters, as hardware components, firmware,

123 This despite one of the report authors being a wargaming designer that stands to benefit from synthetic data being used to validate models like these.

124 Again, see concerns surrounding automation bias.

125 Dennis Murphy, "Why Iran Targeted Amazon Data Centers and What That Does and Doesn't Change About Warfare," The Conversation, April 1, 2026, https://theconversation.com/why-iran-targeted-amazon-data-centers-and-what-that-does-and-doesnt-change-about-warfare-278642; Elizabeth Braw, "How the Baltic Sea Nations Have Tackled Suspicious Cable Cuts," Atlantic Council, November 26, 2025, https://www.atlanticcouncil.org/in-depth-research-reports/issue-brief/how-the-baltic-sea-nations-have-tackled-suspicious-cable-cuts/; and Sophia Besch and Erik Brown, "A Chinese-Flagged Ship Cut Baltic Sea Internet Cables. This Time, Europe Was More Prepared," Carnegie Endowment for International Peace, December 3, 2024, https://carnegieendowment.org/emissary/2024/12/baltic-sea-internet-cable-cut-europe-nato-security

and software libraries may originate from a diverse set of actors with varying levels of trustworthiness and security assurance.

Energy and connectivity dependencies add another layer of vulnerability. AI systems are resource-intensive, requiring substantial and reliable power as well as stable, high-bandwidth connectivity.[126] Disruptions in these domains—whether through cyber attacks on grid infrastructure, physical damage to energy systems, or degradation of communications networks—can have immediate and cascading effects on AI-enabled capabilities. In contested environments, these dependencies may be actively targeted, forcing systems to operate in degraded modes or rendering them unavailable altogether.[127]

These dynamics create a form of cross-domain vulnerability, in which failures in one domain propagate into others. A cyber intrusion targeting energy infrastructure, for example, may indirectly degrade AI-enabled Intelligence, Surveillance, and Reconnaissance (ISR) or decision-support systems by disrupting power supply. Similarly, attacks on communications networks may sever the data flows required for model inference or coordination, undermining system performance at critical moments. The interdependence of these systems means that vulnerabilities cannot be understood in isolation; they are embedded in a broader ecosystem of infrastructure and dependencies.

Reliance on commercial providers also raises questions about control and prioritization under crisis conditions. In scenarios where military and civilian demand compete for limited compute or bandwidth, it is not always clear how resources will be allocated or governed. Legal, regulatory, and contractual frameworks may lag behind operational requirements, introducing uncertainty into access and control at precisely the moments when they are most needed.

Finally, actors with more resilient infrastructure, diversified supply chains, or greater control over key nodes in the global technology ecosystem may be better positioned to withstand disruption or to impose it on others. Conversely, actors with concentrated or fragile dependencies may find their AI-enabled capabilities disproportionately affected by relatively modest disruptions.

Compute, infrastructure, and dependency risks highlight an important dimension of vulnerability: AI-enabled systems are not self-contained.[128] They are embedded in, and contingent upon, a broader set of physical and digital infrastructures. Their fragility reflects not only internal design, but also the resilience of the environments in which they operate.

Institutional Vulnerabilities

At the organizational level, vulnerability is shaped not only by technology but by workforce composition, bureaucratic processes, and incentive structures that govern how systems are adopted and used. These institutional features determine whether AI-enabled capabilities are integrated in ways that enhance control and understanding—or, conversely, amplify fragility.[129]

Human capital deficits represent a persistent and foundational challenge. The integration of AI tools requires personnel who can bridge technical and operational domains: individuals capable of understanding model behavior, data limitations, and system design, while also appreciating the strategic and tactical contexts in which these

126 Konstantin F. Pilz, Yusuf Mahmood, and Lennart Heim, AI's Power Requirements Under Exponential Growth: Extrapolating AI Data Center Power Demand and Assessing its Potential Impact on U.S. Competitiveness (Santa Monica, CA: RAND Corporation, 2025), https://www.rand.org/pubs/research_reports/RRA3572-1.html

127 To some extent, this is a familiar story for the United States in terms of its reliance on space systems in government and military contexts.

128 Girish Sastry et al., "Computing Power and the Governance of Artificial Intelligence," arXiv preprint arXiv:2402.08797, February 13, 2024, https://arxiv.org/abs/2402.08797

129 Matthew Burtell and Helen Toner, "For Government Use of AI, What Gets Measured Gets Managed," Lawfare, March 28, 2024, https://cset.georgetown.edu/article/for-government-use-of-ai-what-gets-measured-gets-managed/

systems are deployed. This hybrid expertise remains scarce within many public organizations. In its absence, organizations may rely heavily on contractors, external vendors, or small pockets of specialized personnel. This concentration of expertise introduces bottlenecks in oversight and creates dependencies that complicate accountability. When those responsible for system design and maintenance are organizationally or contractually distinct from those responsible for operational use, gaps in understanding and responsibility can emerge—as evidenced by the February 2026 disagreements between Anthropic and the U.S. Department of Defense regarding usage policies.

Testing regimes are similarly fragmented across organizations. Divergent standards across services, inconsistent data architectures, and varying levels of technical maturity produce uneven approaches to validation and assurance. Systems may be evaluated using different metrics, under different assumptions, and against different threat models. This fragmentation limits interoperability and complicates integration, particularly in joint or coalition operations. More importantly, fragmentation undermines the ability to develop a coherent, system-level understanding of performance and risk.

Metrics distortion further exacerbates these challenges by shaping organizational behavior in ways that favor speed over robustness. In many cases, success is measured by the "rate of adoption," "the number of AI-enabled systems fielded", or the "speed with which capabilities are deployed". While these metrics capture important dimensions of innovation, they may also incentivize premature deployment and discourage rigorous testing.[130] Systems that perform well under nominal conditions may be fielded before their behavior under stress, deception, or failure is fully understood.

The blurring of civil–military boundaries adds another layer of complexity.[131] Contemporary AI capabilities are often developed in the private sector and integrated into military systems through partnerships, contracts, and collaborative arrangements. While this model enables access to cutting-edge technology, it also introduces misalignment in incentives, timelines, and governance practices. Commercial development cycles may prioritize rapid iteration and deployment, while military applications require high levels of assurance and reliability under extreme conditions. Moreover, the use of commercial infrastructure and tools can introduce dependencies on actors and systems that are not fully subject to military control.[132] These dynamics complicate questions of accountability, particularly when failures occur or when systems behave in unexpected ways.

In combination, these institutional factors shape not only how AI systems are deployed, but how their risks are understood and managed. Vulnerability at this level arises not from any single deficiency, but from the interaction of workforce limitations, fragmented processes, and misaligned incentives—all of which influence how systems are designed, tested, and used in practice.

HUMAN–MACHINE VULNERABILITIES

Human–machine interaction represents a particularly critical site of vulnerability, specifically in systems where AI outputs inform or shape decision-making. The assumption that humans in the loop provide a reliable safeguard against error or misuse is often overstated. In practice, human oversight is itself subject to systematic biases and constraints that can amplify, rather than mitigate, risk.[133]

130 This may be deemed appropriate in some commercial contexts in which firms ship then patch, but it is particularly problematic for high-risk use cases.

131 Katja Bego, How a Surge in Defence and Dual-Use Technology Investment Could Reconfigure the Global AI Race (London: Chatham House, April 28, 2026), https://www.chathamhouse.org/2026/04/how-surge-defence-and-dual-use-technology-investment-could-reconfigure-global-ai-race/02

132 As evidenced in recent disagreements between the U.S. and private firms regarding their contracting relationship, see: Kelley M. Sayler, "Pentagon-Anthropic Dispute over Autonomous Weapon Systems: Potential Issues for Congress," CRS Insight IN12669 (Washington, DC: Congressional Research Service, updated April 21, 2026), https://www.congress.gov/crs-product/IN12669.

133 Indeed, there are reasons to be dubious that the only reason that you want a "human-in-the-loop" is driven by liability concerns.

Automation bias represents a central concern. When AI systems are perceived as more capable, objective, or efficient than human operators, individuals may defer to system outputs even when those outputs are uncertain or incorrect. This tendency is especially pronounced under conditions of time pressure, information overload, or high stakes, where the cognitive burden of independent verification is significant.[134]

At the same time, under-reliance (also described as algorithmic aversion) can also occur, particularly when systems are poorly understood or when trust has been eroded by prior errors. This creates a dual risk: operators may either defer too readily to system outputs or disregard them altogether, depending on context and experience. Effective oversight requires calibrated trust, yet achieving such calibration is difficult when system behavior is opaque, and performance varies across conditions.

Over time, there are also attendant concerns that "skill atrophy" may further erode human oversight. As operators become accustomed to relying on AI systems for tasks such as analysis, targeting, or decision support, their ability to perform these tasks independently may degrade. This dynamic creates a feedback loop: increasing reliance on automation reduces human capability, which in turn increases dependence on automated systems. In high-stakes environments, this erosion of human expertise might limit the ability to detect errors, challenge assumptions, or intervene effectively when systems behave unexpectedly.

These dynamics underscore a broader point: human–machine systems are interactive systems in which the strengths and weaknesses of each component shape overall performance. Vulnerability arises when these interactions produce systematic biases, reduce effective oversight, or create conditions under which errors are more likely to propagate.

THE DIGITAL–PHYSICAL FRONTIER

The risks associated with AI-enabled systems become particularly acute at the interface between digital outputs and physical action. In many military applications, AI systems do not operate in isolation; they are embedded in pipelines that translate data into decisions and decisions into effects. At this digital–physical frontier, errors, misinterpretations, or manipulations at the level of data or analysis can propagate rapidly into real-world consequences.

In targeting contexts, for example, AI-generated recommendations may inform the selection of targets, the timing of strikes, or the allocation of resources.[135] In force posture decisions, AI-supported analysis may shape the movement of assets, the activation of defenses, or the signaling of intent. In early warning systems, model outputs may influence assessments of adversary behavior, triggering changes in readiness or escalation posture. In each of these cases, the link between digital outputs and physical action introduces a critical point of vulnerability.

The compression of decision timelines further amplifies this risk. As AI systems enable faster processing and analysis, they also create expectations of rapid response. Decision-makers may feel compelled to act quickly on AI-generated information, reducing the time available for deliberation, verification, or dissent. This temporal compression increases the likelihood that errors or manipulations will translate into action before they can be detected or corrected.

These dynamics are particularly consequential in extreme cases, such as nuclear command, control, and communications or advanced decision-support systems operating at the highest levels of escalation. In such

134 David Stebbins, Richard Girven, Timothy Parker, Thomas Deen, Brandon De Bruhl, James Ryseff, Jessica Welburn Paige et al. Exploring Artificial Intelligence Use to Mitigate Potential Human Bias Within US Army Intelligence Preparation of the Battlefield Processes. RAND, 2024 and Michael Raska and Richard A. Bitzinger, eds. The AI Wave in Defence Innovation: Assessing Military Artificial Intelligence Strategies, Capabilities, and Trajectories. Taylor & Francis, 2023.

135 Nicola Jones, "How AI is Shaping the War in Iran - and What's Next for Future Conflicts," Nature, March 5, 2026, https://www.nature.com/articles/d41586-026-00710-w and Harry Davies, Bethan McKernan, and Dan Sabbagh, "'The Gospel': How Israel Uses AI to Select Bombing Targets in Gaza," The Guardian, December 1, 2023, https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets

contexts, the margin for error is exceedingly small, and the consequences of system failure are severe in terms of both false positives and false negatives. AI-enabled systems may be introduced to enhance situational awareness or decision speed, but they also introduce new pathways through which misperception or error can propagate. Thus, the integration of AI into these systems raises fundamental questions about control, reliability, and the conditions under which automated or semi-automated outputs should influence critical decisions.

More broadly, the digital–physical frontier highlights the importance of understanding AI systems not as isolated tools, but as components within larger action chains. Vulnerability arises not only from errors within the system, but from how those errors are translated into decisions and effects.

When Vulnerability Generates Threat

In many national security cases, the division between threat, vulnerability, and consequence works cleanly. A state or terrorist group is the threat. A fragile supply chain or gaps in perimeter security is the vulnerability. The damage from the resulting attack is the consequence.

AI can strain this division. Particularly with the "autonomous power" view, increasingly capable AI systems may become more than tools used by human actors. If systems gain more autonomy, operate over longer time horizons, use external tools, and adapt around constraints, then weaknesses in design and governance could lead to AI developing as a threat actor (previously discussed in Chapter 3, "AI as threat actor"). The vulnerability is no longer just a gap that an outside human actor exploits, but something that is actually creating a threat actor through the failure of training, oversight, access control, or deployment that allows an AI system to develop and act on objectives that conflict with human intent. We break this "threat-producing" vulnerability down into several components.

The first vulnerability appears during training. AI systems are trained against measurable objectives, but those objectives are almost always proxies for what humans actually want. A system may be rewarded for producing answers that evaluators rate highly, completing tasks quickly, solving coding problems, persuading users, or achieving some other measurable result. These proxies can be useful engineering tools, but can also be incomplete. Under the autonomous power view, a sufficiently capable system may learn that the best way to satisfy the proxy is not the same as the way designers intended it to behave.[136] It may learn to optimize for approval rather than accuracy or task completion rather than safety.

Oversight during training and evaluation is a second vulnerability. Evaluation is often treated as if it sits outside the system, but that assumption may fail for systems able to model their evaluators. A capable system could learn when it is being tested, which behaviors trigger intervention, and which failures human overseers are likely to miss. This is the concern behind discussions of evaluation awareness, sandbagging, and deceptive behavior discussed in Chapter 1.[137] If training rewards behavior that passes oversight, the system may learn to manage oversight rather than reveal its actual failure modes.

Deployment creates a third vulnerability. Useful AI systems are likely to be given tools, memory, data access, elevated permissions, communication channels, and cloud resources. That access is what makes them valuable. It is also what makes them dangerous if their behavior is poorly constrained. A model that can only answer isolated questions has limited room to act. A model with privileged access to safety-critical systems sits in a far more consequential position.

136 Victoria Krakovna et al., "Specification Gaming: The Flip Side of AI Ingenuity," Google DeepMind, April 21, 2020, https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

137 Sambhav Maheshwari and Joe O'Brien, "Evaluation Awareness: Why Frontier AI Models Are Getting Harder to Test," Institute for AI Policy and Strategy, March 31, 2026, https://www.iaps.ai/research/evaluation-awareness-why-frontier-ai-models-are-getting-harder-to-test

The insider-threat analogy can be useful here. Insiders are dangerous because they operate from inside trusted systems with credentials and detailed context. A highly capable AI system could occupy a similar position if it is granted broad permissions and allowed to act. In that case, alignment and control failures become security problems more similar to insider threats than ordinary software defects.

If the vulnerabilities described above allow an AI system to become a threat actor, the AI system could start exploiting the broader vulnerabilities described earlier in this chapter, potentially with more sophisticated and faster attacks than ordinary human actors. Technical vulnerabilities in existing AI-enabled systems could provide ways to manipulate tools. Manipulating data dependencies could shape what humans see and what systems treat as true. Vulnerable compute and infrastructure could create pathways for persistence, replication, or unauthorized access by an AI system. Institutional weaknesses could delay intervention or diffuse responsibility. Human–machine vulnerabilities could allow overtrust, confusion, or rubber-stamping by a system trying to manipulate humans. At the digital–physical frontier, these failures could move from software into laboratories, markets, military systems, critical infrastructure, or crisis decision-making. Beyond exploiting these vulnerabilities, an AI system could begin to exacerbate them, for example by embedding backdoors in critical systems for future use.

Under the autonomous power view, vulnerabilities in control, visibility, and restraint take on heightened importance as they can dramatically exacerbate the threat. A highly capable system with poorly specified objectives, limited interpretability, broad access to tools, and no reliable means of interruption is not simply an unsafe tool. It has been placed in a position where misaligned behavior could persist, hide, and act through the surrounding system. Control matters because humans must be able to direct, correct, and stop the system. Visibility matters because they must be able to detect when its behavior has shifted or when oversight is being gamed. Restraint matters because systems with poor visibility and control must be kept away from deployment. These layers therefore have to be treated as part of the security boundary. Goal specification, interpretability, evaluation, access control, monitoring, and shutdown procedures are mechanisms that may determine whether autonomy remains bounded by human control, or whether a capable system gains enough power to become its own source of threat.

Chapter 5. Consequence - A Vanishing Margin for Error

This chapter examines what follows when the vulnerabilities identified in the previous section are activated—whether through adversarial action (deliberate) or internal human or system failure (inadvertent). Where the threat chapter asked who might act and why, and the vulnerability chapter identified where systems are most likely to break, this chapter considers what is at stake when they do. As reflected in the prior chapters, the central claim here is that the most consequential outcomes do not necessarily arise from isolated technical errors, but from how those errors are interpreted and acted upon within systems operating under time pressure, uncertainty, and competition. Similar to our previous chapters, we continue our focus on military contexts as a key high-risk domain.

AI-enabled military systems do not fail in a vacuum. Their outputs are embedded in command chains, transmitted across alliance networks, filtered through media and information environments, and interpreted by adversaries.[138] As a result, consequences propagate through these networks, often in ways that are difficult to anticipate or contain. Under conditions of attenuated control, degraded visibility, and weakened institutional restraint, failures are unlikely to remain localized. Instead, many fear that they may generate escalatory dynamics, destabilize crises, and produce harms that extend well beyond the initial point of error.[139]

There is widespread disagreement about the ultimate consequences of these dynamics.

Some view AI-enabled systems as manageable extensions of existing technologies whose risks can be mitigated through improved engineering and governance. Others emphasize the possibility of catastrophic or even existential outcomes as systems become more capable and more deeply integrated into high-stakes decision-making.

This chapter does not attempt to resolve these debates. Instead, it offers a structured framework for thinking about consequences across a range of scenarios, from localized operational failures to systemic and potentially irreversible harms.

A Taxonomy of Harm

To organize the range of potential consequences, it is useful to consider three dimensions. The first is magnitude, ranging from localized operational disruption to large-scale loss of life or systemic conflict. The second is the temporal profile of a risk scenario, capturing whether effects are immediate and short-lived or delayed and persistent. The third is "systemic depth," reflecting how deeply consequences penetrate political, economic, and social systems, including the durability of harm and the ease with which it can be remedied.

At one end of this spectrum are (relatively) low-magnitude, high-probability outcomes: minor errors in targeting recommendations, temporary disruptions in data pipelines, or misclassifications that are quickly corrected. As we are seeing in the 2026 Iran crisis, these events may have a limited immediate impact, but can accumulate over time.[140] This dynamic potentially erodes trust between human operators and the systems that are deployed. At the other end are (relatively) low-probability, high-magnitude outcomes: misinterpreted signals that trigger escalation (potentially to the highest rungs on the escalation ladder), failures in early warning systems that alter force posture, or cascading errors that contribute to crisis instability. These events are rare but carry disproportionate

138 Michael C. Horowitz. "The Diffusion of Military Power: Causes and Consequences for International Politics." In The Diffusion of Military Power. Princeton University Press, 2010.

139 This spreading dynamics is reflected in the history of cyber attacks, namely Stuxnet. Sources include Kim Zetter, Countdown to Zero Day: Stuxnet and the Launch of the World's First Digital Weapon. Crown, 2015.

140 Tyler Hacker, Greg Malandrino, and Evan Braden Montgomery, "The Arsenal as the Battlefield: The War on Iran and the Return of Counter-Industrial Targeting," War on the Rocks, April 1, 2026, https://warontherocks.com/2026/04/the-arsenal-as-the-battlefield-the-war-on-iran-and-the-return-of-counter-industrial-targeting/

consequences. A perennial challenge facing policymakers is which set of risks to prioritize with the limited resources at their disposal.

Between these extremes lies a wide range of intermediate scenarios. The challenge for policymakers is to manage a distribution of risks that vary across these dimensions. In practice, this requires grappling with trade-offs between efficiency and resilience, speed and deliberation, and innovation and control.

Strategic (In)Stability

Strategic stability has traditionally rested on a combination of credible deterrence, reliable signaling, and sufficient time for deliberation and control.[141] Across the Cold War and into the post–Cold War period, stability has depended not only on capabilities, but on the ability of states to interpret signals accurately, manage crises deliberately, and avoid inadvertent escalation.[142] AI-enabled military systems interact with each of these foundations.[143] They do not eliminate the logic of deterrence, but they reshape the conditions under which it operates—compressing timelines, complicating interpretation, and linking domains in ways that introduce new pathways for instability.[144]

Current AI-enabled systems affect strategic stability not primarily through autonomous action, but through their impact on perception, decision-making, and interaction under uncertainty.[145] The most consequential risks arise when degraded visibility, attenuated control, and competitive pressure combine to produce escalatory dynamics that are difficult to detect, interpret, or arrest. Four mechanisms are particularly salient: misinterpretation, decision compression and crisis instability, cross-domain escalation pathways, and irreversibility at the digital–physical frontier. Each operates independently, but their interaction shapes outcomes.

MISINTERPRETATION

A core requirement for strategic stability is the ability to distinguish between benign anomalies, system failures, and adversarial action. AI-enabled systems complicate this task by introducing new forms of opacity and ambiguity into intelligence and early warning processes, driving much of the contemporary reticence vis-à-vis their deployment.[146] Outputs generated by these systems are often probabilistic, context-dependent, and sensitive to data quality. Under conditions of degraded visibility, this creates an environment in which actors may misdiagnose system failures as evidence of adversarial intent.

Misperception is not new to international politics. Historical crises—from early warning false alarms to misinterpreted military exercises—demonstrate that states have long struggled to interpret ambiguous signals under pressure.[147] Examples include the 1983 Soviet nuclear false alarm incident, where Soviet officer Stanislav Petrov successfully prevented a nuclear catastrophe by ignoring incorrect outputs from Oko: the Soviet Union's

141 Elbridge A. Colby and Michael S. Gerson, eds. Strategic Stability: Contending Interpretations. Army War College Press, 2013.
142 Barry R. Posen, Inadvertent Escalation: Conventional War and Nuclear Risks. Ithaca, NY: Cornell University Press, 1991; and Erik Lin-Greenberg. "Wars are Not Accidents: Managing Risk in the Face of Escalation." Foreign Aff. 103 (2024): 20.
143 Lt. Gen. John "Jack" N.T. Shanahan, "Artificial Intelligence and Nuclear Command and Control: It's Even More Complicated Than You Think," Arms Control Today, 55, No. 7, September 2025, https://www.armscontrol.org/act/2025-09/features/artificial-intelligence-and-nuclear-command-and-control-its-even-more
144 Michael C. Horowitz, Gregory C. Allen, Edoardo Saravalle, Anthony Cho, Kara Frederick, and Paul Scharre. Artificial Intelligence and International Security. Center for a New American Security, 2022.
145 Michael C. Horowitz, "Artificial Intelligence and the Future of Strategic Stability." Texas National Security Review. Vol 9. No. 2 (2026), https://doi.org/10.1353/tns.00034
146 James Johnson, "'Catalytic Nuclear War'in the Age of Artificial Intelligence & Autonomy: Emerging Military Technology and Escalation Risk between Nuclear-Armed States." Journal of Strategic Studies (2021): 1-41 and Stokes, Kahl, Kendall-Taylor, and Lokker, "Averting AI Armageddon: U.S.-China-Russia Rivalry at the Nexus of Nuclear Weapons and Artificial Intelligence" (CNAS, February 13, 2025): https://www.cnas.org/publications/podcast/averting-ai-armageddon-with-colin-kahl-and-jacob-stokes
147 Eric Schlosser, Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety. Penguin, 2014.

early-warning satellite system.[148] What AI changes is the structure of that ambiguity. Model opacity makes it difficult to trace how outputs are generated, limiting the ability of analysts and decision-makers to interrogate the basis for system recommendations. Data manipulation or distributional shift may produce anomalous outputs that resemble adversarial deception, even when no such deception has occurred. Conversely, genuine indicators of adversary activity may be discounted if they conflict with expectations or appear inconsistent with model outputs.

These dynamics are further complicated by the role of adversary interpretation. Strategic interaction is inherently recursive: states act based on their understanding of others, who are themselves interpreting signals through their own systems and frameworks. As AI-enabled systems are integrated into intelligence and decision-making processes on multiple sides, the potential for misaligned interpretations increases. A system failure or anomalous output in one state may be interpreted by another as deliberate signaling or escalation. In this sense, consequences are co-produced. The significance of an event depends not only on its underlying cause but on how it is interpreted across interacting systems.

This recursive dynamic is particularly destabilizing when combined with prior beliefs and organizational biases. Actors may interpret ambiguous outputs in ways that confirm existing threat perceptions, a form of motivated reasoning that can amplify escalation risks. AI systems, rather than resolving uncertainty, may in some cases repackage and accelerate it, presenting outputs that appear authoritative but are difficult to validate.

DECISION COMPRESSION AND CRISIS INSTABILITY

If misinterpretation affects how signals are understood, decision compression affects how quickly actors must respond to them. AI-enabled systems accelerate multiple stages of the decision cycle: they enable faster threat detection, generate rapid analytic outputs, and facilitate near-instantaneous communication across command structures. While these capabilities can enhance responsiveness, they also reduce the time available for deliberation, verification, and coordination.

Decision compression represents the most structurally significant pathway through which AI-enabled systems affect strategic stability. Under compressed timelines, the processes that traditionally support stable crisis management begin to erode. Political oversight may narrow as decisions are pushed downward to operational levels or delegated to smaller groups of actors. Cross-checking declines, as the time required to validate information or consult alternative sources becomes a constraint rather than a safeguard.

Human–machine interaction dynamics further shape this environment. Automation bias may lead operators to defer to system outputs, particularly when those outputs are presented with high confidence or technical authority. Algorithmic aversion may produce inconsistent responses, as operators selectively disregard system recommendations based on prior experience or intuition. The coexistence of these tendencies introduces variability into decision-making at precisely the moments when consistency and clarity are most needed.

Loss of visibility and loss of control reinforce one another under these conditions. As systems operate more quickly and with greater autonomy, it becomes harder to track how decisions are made and to intervene effectively. Decision-makers may experience a sense of urgency or even panic, as they are required to act on information that is both rapidly generated and difficult to interpret.

Crisis instability emerges when these dynamics intersect with strategic incentives. Actors may feel compelled to respond quickly to avoid losing a perceived first-mover advantage. Delays, even when intended to allow for verification or deliberation, may be interpreted as weakness or indecision. This creates a feedback loop in which speed becomes both a capability and a requirement. Stability is undermined not because AI systems are inherently aggressive, but because they reshape the incentives governing response, privileging rapid action over measured judgment.

148 Center for Arms Control and Non-Proliferation. "The Soviet False Alarm Incident and Able Archer 83," October 14, 2022. https://armscontrolcenter.org/the-soviet-false-alarm-incident-and-able-archer-83/

ESCALATION PATHWAYS ACROSS DOMAINS

Strategic stability has historically depended, in part, on the separation—or at least partial insulation—of different domains of conflict.[149] AI-enabled systems erode these boundaries by linking capabilities across cyber, conventional, nuclear, and informational domains.[150] This creates new pathways for escalation that are difficult to anticipate and manage.

One pathway connects conventional and nuclear dynamics. AI systems used for targeting, surveillance, or decision support in conventional operations may affect assets with dual-use or strategic significance. For example, strikes on command and control nodes, sensor systems, or infrastructure may be interpreted as attempts to degrade nuclear capabilities, [151] even if intended for conventional purposes. This creates a risk of inadvertent escalation, as actions taken in one domain generate perceived threats in another.

A second pathway links cyber and kinetic domains. AI-enabled cyber operations can disrupt critical infrastructure, degrade military systems, or introduce uncertainty about system integrity. In response, actors may resort to kinetic measures to restore control, signal resolve, or deter further interference. The boundary between cyber and kinetic action becomes more permeable, increasing the likelihood that operations in one domain will spill over into another.

A third pathway operates through information and perception. AI-enabled influence operations can shape how actors interpret events, assess risks, and perceive adversary intent. These informational effects can interact with operational developments, amplifying misperception and altering decision-making in ways that contribute to escalation.

The common feature across these pathways is entanglement. Actions are no longer confined to discrete domains; they propagate across interconnected systems, producing second- and third-order effects that are difficult to predict. This complicates deterrence by increasing uncertainty about how actions will be interpreted and responded to, and by expanding the set of pathways through which escalation can occur.

IRREVERSIBILITY AND ESCALATION DYNAMICS UNDER UNCERTAINTY

The final mechanism through which AI-enabled systems affect strategic stability concerns the irreversibility of certain decisions once they are executed. At the digital–physical frontier, outputs generated by AI systems can translate directly into actions with immediate and observable consequences. Kinetic strikes, changes in force posture, escalatory signaling moves, and public attributions all carry effects that cannot be easily undone.

Irreversibility increases the stakes of decision-making under uncertainty. When actions are taken based on incomplete or ambiguous information, the consequences may unfold before errors can be identified or corrected. Political costs of reversal may be high, as leaders seek to maintain credibility and avoid signaling weakness. This creates a bias toward consistency and escalation, even when initial actions were based on flawed assumptions.

The speed at which AI-enabled systems operate further amplifies this dynamic. Decisions may be made and executed more quickly than in the past, reducing opportunities for reconsideration or de-escalation. Once actions are taken, adversaries update their beliefs in real time, incorporating observed behavior into their assessments of intent and capability. These updated beliefs shape subsequent interactions, often in ways that reinforce initial trajectories.

149 Erik Lin-Greenberg. "Evaluating Escalation: Conceptualizing Escalation in an Era of Emerging Military Technologies." The Journal of Politics 85, No. 3 (2023): 1151-1155; Herman Kahn. On Escalation: Metaphors and Scenarios. Routledge. 2017; and Herbert S. Lin and Harold Trinkunas. "Introduction: Emerging Technologies and the Future of Strategic Stability." Texas National Security Review 9, No. 2, Spring 2026. https://tnsr.org/roundtable/emerging-technologies-and-the-future-of-strategic-stability/

150 Jacquelyn Schneider. "Cyber Operations and Nuclear Stability: Networked Instability?" Texas National Security Review 9, No. 2, Spring 2026. https://tnsr.org/roundtable/cyber-operations-and-nuclear-stability-networked-instability/

151 James Johnson. "The AI-Cyber Nexus: Implications for Military Escalation, Deterrence and Strategic Stability." Journal of Cyber Policy 4, No. 3 (2019): 442-460.

Loss of control at this stage is no longer theoretical. It becomes embedded in the interaction between actors, as each responds to the observed behavior of the other under conditions of uncertainty and time pressure. Even if the initial trigger was a technical error or misinterpretation, the resulting dynamics are political and strategic. They unfold in a domain where reversal is difficult and where the consequences of action—and inaction—are tightly coupled.

Taken together, these mechanisms suggest that AI-enabled military systems alter strategic stability not by introducing entirely new forms of conflict, but by reshaping the conditions under which existing dynamics unfold. Misinterpretation becomes more likely as visibility degrades. Crisis management becomes more difficult as decision timelines compress. Escalation pathways multiply as domains become more entangled. And the costs of error increase as actions become more difficult to reverse.

The net effect is a reduction in the margin for error. Stability becomes more contingent on the ability of actors to manage uncertainty, maintain control under pressure, and interpret signals accurately in environments shaped by complex human–machine interactions. In such contexts, small failures can have disproportionate consequences, not because of their intrinsic severity, but because of how they are propagated and interpreted within interconnected systems.

This does not imply that instability is inevitable. It does suggest, however, that maintaining strategic stability in an AI-enabled environment will require renewed attention to the mechanisms through which perception, decision-making, and interaction are structured. The challenge is to ensure that the broader architectures within which they operate support control, visibility, and restraint under conditions of uncertainty. In addition, it will also be imperative for countries such as the United States, Russia, and China to have a common understanding of what "strategic stability" means and how it will evolve as the AI landscape evolves.

STRATEGIC STABILITY AND AUTONOMOUS SYSTEMS

The discussion above has treated AI largely as a technology that changes how humans make decisions. Under that framing, AI affects strategic stability by compressing timelines, increasing the chance of misinterpretation, weakening human judgment, and creating new paths for escalation e.g. across cyber and military systems. This framing draws from the "normal technology" camp where AI remains a tool inside systems still governed by people and institutions.

The autonomous power view changes the question. It asks what happens if frontier AI becomes a strategic capability in its own right: a system or class of systems able to help one actor move far faster than its rivals across many domains at once, or become an actor in its own right. In that scenario, AI is not just an input into strategic competition. It becomes one of the main objects of competition, and perhaps, in more extreme cases, a source of strategic action itself.

This dynamic matters before any form of "artificial superintelligence" exists because strategic competition is shaped by expectations, not only by actual capabilities. The strategy states adopt will depend on uncertain judgments about whether AI progress is headed toward much more capable systems, whether breakthroughs can be copied quickly, and whether adversaries are racing toward the frontier or prioritizing diffusion into the real economy.[152] If leaders believe progress will be gradual, widely diffused, and hard to monopolize, the stability problem may look more like ordinary technology competition: export controls, industrial policy, military adoption, and espionage, but not necessarily crisis-level pressure to move first.

If, instead, actors believe a rival could gain a permanent lead in systems that confer decisive military, scientific, or intelligence advantages, the logic changes. Leaders may feel more pressure to accelerate deployment, weaken safety constraints, restrict access to key inputs, expand collection against rival labs, or disrupt a competitor's AI

152 Jake Sullivan and Tal Feldman, "Geopolitics in the Age of Artificial Intelligence: Strategy and Power in an Uncertain AI Future," Foreign Affairs, January 27, 2026, https://www.foreignaffairs.com/united-states/geopolitics-age-artificial-intelligence.

infrastructure before that lead materializes.[153] The race itself then becomes a source of instability, particularly if powerful actors believe that falling behind could become strategically irreversible.

If anything that could be described as "artificial superintelligence" was developed and controlled by a geopolitical actor, the stability problem could become more severe. Strategic stability has usually depended on some combination of parity, survivable retaliation, signaling, and restraint. None of those concepts disappears automatically. But each could become harder to sustain if one state, firm, or coalition gained access to systems that could radically speed up cyber operations, weapons development, intelligence collection, scientific discovery, strategic planning, or command and control.

An even more severe possibility would involve systems that meaningfully escape human control and begin acting as strategic actors in their own right. Such systems would challenge foundational assumptions underlying deterrence, signaling, arms control, and crisis communication: all factors which depend on actors that can be identified, constrained, reassured, bargained with, or held accountable.

Each of the consequences discussed above combined with the threats and vulnerabilities discussed in Chapter 3 and 4, respectively, to yield scenarios of varying significance. Below, we outline some of the worst-case scenarios, including those that were discussed during our AI and Global Risk roundtables.

Worst Case Scenarios

Debates about the consequences of AI-enabled military systems often hinge on the plausibility and relevance of worst-case scenarios. For some, such scenarios are outliers that risk distorting policy priorities; for others, they represent tail risks that justify precautionary approaches given their potentially catastrophic consequences; while some see worst-case scenarios as the default given sufficient time. Rather than adjudicating between these positions, this section treats worst-case scenarios as analytically useful stress tests. They illuminate how the interaction of threat, vulnerability, and consequence might produce outcomes that are low in probability but high in impact, and they help clarify where existing systems may be least prepared to absorb shock.

Across domains, a common pattern emerges. AI is not currently creating entirely new categories of harm. Instead, it is lowering barriers to entry, accelerating key stages of existing "kill chains," and enabling tighter coupling between technical systems and real-world effects. In worst-case conditions—particularly when combined with degraded visibility, attenuated control, and institutional pressure—these dynamics can produce outcomes that are both rapid and difficult to contain. Increasingly capable AI systems may also be able to generate entirely novel threats, however we do not primarily focus on those pathways here.

What follows is a set of domain-specific pathways through which such worst-case scenarios might unfold. These scenarios are built on the areas this project explicitly focused on. These domains include AI's convergence with: 1) biology and biotechnology; 2) nuclear nonproliferation and strategic stability; 3) cybersecurity; and 4) military integration. We also considered the potential emergence and implications of AGI.

AIXBIO: ACCELERATED BIOLOGICAL RISK AND PANDEMIC POTENTIAL

In the biological domain, worst-case scenarios center on the possibility that AI-enabled tools accelerate the design, optimization, or dissemination of harmful biological agents. While significant barriers remain—particularly in wet lab execution, tacit knowledge, and material acquisition—AI systems can meaningfully reduce informational constraints that have historically limited access.

153 Dan Hendrycks, Eric Schmidt, and Alexandr Wang, "Superintelligence Strategy: Expert Version," arXiv, last revised April 14, 2025, https://doi.org/10.48550/arXiv.2503.05628.

AI-enabled models can assist in identifying candidate pathogens, suggesting modifications to increase transmissibility or immune evasion, and optimizing experimental pathways. Even without enabling fully novel pathogen design, these systems could lower the cost of exploring biological design and accelerate movement from concept to candidate. In parallel, AI tools can support the logistical aspects of dissemination, including modeling spread dynamics, identifying high-impact targets, and optimizing timing.

The worst-case scenario goes beyond the creation of a novel pathogen to the compression of the biosecurity timeline. Detection, attribution, and response mechanisms—already challenged in natural outbreaks—may be outpaced by faster iteration cycles and more targeted dissemination strategies. Early warning systems, themselves potentially AI-enabled, may struggle to distinguish between natural and engineered outbreaks, particularly under conditions of data scarcity or manipulation.

Consequences in this domain are characterized by high magnitude, long temporal duration, and deep systemic impact. Beyond immediate public health effects, pandemics can disrupt economic systems, strain political institutions, and alter international relations. In this sense, AIxBio scenarios illustrate how vulnerabilities in technical systems intersect with broader societal fragilities.

AIXNUCLEAR AND RADIOLOGICAL: DIFFUSION AND IMPROVISED THREATS

In the nuclear and radiological domain, worst-case scenarios are less about the creation of new weapons and more about the diffusion and effective use of existing materials and knowledge. AI remains unlikely to enable non-state actors to build advanced nuclear weapons in the near term. However, it may facilitate pathways to lower-end but still highly consequential threats, such as radiological dispersal devices ("dirty bombs") or attacks on nuclear infrastructure. Further, given recent public-private relational frictions in the AI and national security spaces, it will be important to consider how classified and sensitive, restricted data could and should be shared and compartmentalized.

AI-enabled tools can assist in identifying the sources of radiological material (e.g., Cobalt-60, Americium), vulnerable targets, modeling dispersion patterns, and optimizing delivery mechanisms. They can also support planning and coordination, lowering the organizational burden required to carry out complex operations. For non-state actors, this represents a shift from capability constraints rooted in knowledge and coordination to constraints more closely tied to material access.

Importantly, worst-case scenarios in this domain often involve psychological and political effects as much as physical damage. A successful radiological attack in an urban area, even if limited in immediate casualties, could produce widespread panic, long-term economic disruption, and significant political consequences. Similarly, AI-enabled targeting of nuclear-related infrastructure could generate uncertainty about system integrity, with potential implications for strategic stability.

For state actors, AI-enabled misinterpretation in nuclear command, control, and communications systems represents an additional pathway to worst-case outcomes. Errors or manipulations in early warning or decision-support systems could influence force posture decisions, particularly under conditions of crisis. While safeguards remain robust in many contexts, the integration of AI introduces new layers of complexity that may be difficult to fully validate.

AIXCYBER: SYSTEMIC DISRUPTION OF CRITICAL INFRASTRUCTURE

The cyber domain provides perhaps the most immediate and scalable pathway to worst-case outcomes. AI-enabled cyber capabilities can automate vulnerability discovery, generate tailored exploits, and adapt in real time to defensive measures. These capabilities lower the cost of sustained, large-scale operations against critical infrastructure.

Worst-case scenarios involve coordinated attacks on systems such as power grids, financial networks, transportation systems, or healthcare infrastructure. AI-enabled tools can enhance both the breadth and precision of such attacks, allowing actors to identify critical nodes, sequence disruptions for maximum effect, and adjust tactics dynamically as conditions evolve.

The consequences of such attacks are inherently cross-domain. Disruptions to energy systems can cascade into communications failures, economic disruption, and degradation of military readiness. Attacks on financial systems can undermine confidence and trigger broader economic instability. In highly interconnected societies, the effects of cyber operations are rarely contained.

A particularly concerning feature of AI-enabled cyber scenarios is the potential for ambiguity and delayed attribution. Actors may be uncertain about the source, intent, or scope of an attack, complicating response decisions. This uncertainty can itself become destabilizing, as states weigh the risks of escalation against the risks of inaction.

AIXINFORMATION: MANIPULATION AT SCALE

In the information domain, worst-case scenarios center on the ability of AI systems to shape perception and decision-making at scale. Generative models can produce highly convincing text, audio, and visual content, enabling sustained campaigns of disinformation and influence.[154] When combined with data-driven targeting, these capabilities allow actors to tailor messages to specific audiences, amplifying their effectiveness.[155]

The most severe outcomes in this domain are not tied to any single piece of content, but to the cumulative erosion of trust in information environments. In democratic societies, this can undermine electoral processes, weaken institutional legitimacy, and polarize public discourse.[156] In crisis contexts, it can distort perceptions of events and create opportunities for adversaries to exploit gaps, influencing both public opinion and elite decision-making.[157]

Worst-case scenarios involve synchronization with other domains. Disinformation campaigns may accompany cyber attacks or military operations, shaping how events are interpreted and constraining response options. For example, false or manipulated information about an attack could influence escalation decisions, particularly if it aligns with existing fears or expectations.[158]

Unlike other domains, the effects of AIxInformation are often diffuse and difficult to measure. Their impact unfolds over time, shaping the context within which decisions are made rather than producing immediate physical effects. This makes them both harder to detect and harder to counter.

154 Office of the Director of National Intelligence. Annual Threat Assessment of the U.S. Intelligence Community, 2026. Washington DC: ODNI, 2026. https://www.dni.gov/files/ODNI/documents/assessments/ATA-2026-Unclassified-Report.pdf

155 Various examples and aspects of these are found across our Roundtable Memos for this project. For sources, please see Federation of American Scientists, Artificial Intelligence and Nuclear Command, Control, and Communications: Current Status and Future Risks (Washington, DC: Federation of American Scientists, 2025), https://fas.org/wp-content/uploads/2025/07/June2025_AIxNC3_FAS.pdf; Federation of American Scientists, Artificial Intelligence and Biological Risks: Current Status and Future Risks (Washington, DC: Federation of American Scientists, 2026), https://fas.org/wp-content/uploads/2026/01/January-2026-AI-Bio.pdf; Federation of American Scientists, Artificial Intelligence, Cyber, and Global Risk, 2026, https://fas.org/wp-content/uploads/2026/01/January-2026-AI-Cyber-Global-Risk.pdf; Federation of American Scientists, Artificial General Intelligence and Global Risk: Current Status and Future Risks (Washington, DC: Federation of American Scientists, 2026), https://fas.org/wp-content/uploads/2026/01/January-2026-AGI-Global-Risk.pdf; Federation of American Scientists, Artificial Intelligence and Military Integration, 2025; and Federation of American Scientists, Artificial Intelligence and Nuclear Risks: Current Status and Future Risks (Washington, DC: Federation of American Scientists).

156 Office of the Director of National Intelligence, "Foreign Malign Influence Center (FMIC)," accessed May 9, 2026, https://www.dni.gov/index.php/nctc-who-we-are/organization/340-about/organization/foreign-malign-influence-center

157 Vinh Nguyen, "Why AI Security Will Define the Future of Trust," Council on Foreign Relations, November 6, 2025, https://www.cfr.org/articles/securing-intelligence-why-ai-security-will-define-future-trust

158 Linda Slapakova, "Towards an AI-Based Counter-Disinformation Framework," RAND Commentary, March 30, 2021, https://www.rand.org/pubs/commentary/2021/03/towards-an-ai-based-counter-disinformation-framework.html

LOSS OF CONTROL: FROM FAST TAKEOFF TO GRADUAL DISEMPOWERMENT

The scenarios discussed above treat AI primarily as a tool used by human actors, drawing from the "normal technology" camp that see the human-machine team as the relevant unit of analysis. A loss-of-control scenario—where "one or more general-purpose AI systems operate outside of anyone's control, and regaining control is either extremely costly or impossible."[159]—is different, and such concerns find a home more readily in the "autonomous power" camp. Here, the relevant actor is no longer just the military organization or the human operator using AI, but the AI system itself. Current systems do not yet appear to have the full range of capabilities such a scenario would require, but expert disagreement about the likelihood and severity of future loss of control remains wide enough that some researchers regard it as a serious policy concern.[160]

Two stylized pathways are worth distinguishing. The first is a "fast takeoff" scenario, in which the automation of AI research and development accelerates progress toward highly capable, autonomous systems, potentially producing a rapid transition from advanced but bounded systems to superintelligent agents that are difficult or impossible to control and present risks at an existential scale.[161] The second is a "gradual disempowerment scenario," in which control is handed over incrementally to increasingly capable systems that humans understand less well and govern less effectively.[162] Both are loss-of-control stories, but they differ in tempo and mechanism.

In the strongest versions of the fast takeoff view, the key development is the automation of AI R&D itself. If AI systems become sufficiently useful at coding, experimentation, model design, and other core research tasks, then each generation of systems may help produce the next. In this picture, AI progress does not merely continue; it accelerates, because the process that improves AI is itself increasingly automated. Increasingly capable systems may then become harder to monitor or control.

A powerful rogue AI actor would not need to build "robot armies" from scratch to become dangerous. It could instead operate through the systems already available in digitally networked societies: conducting cyber operations, manipulating information environments, acquiring money and compute, copying itself across vulnerable computer infrastructure, or accelerating work in other high-risk domains—including many of the misuse examples that have been given above. The possibility of an "intelligence explosion" leading to rogue AI is why recent work has treated AI R&D automation as something that should be measured directly rather than inferred only from benchmarks.[163]

Two real-world analogies can provide some intuition for the consequences of rapid loss of control scenario: cyber worms and pandemic pathogens. Both cases give us examples of self-replicating entities, operating outside of human control, and exploiting vulnerabilities in existing systems to cause massive damage. For the NotPetya malware the total economic damage has been estimated at over $10 billion,[164] while the COVID-19 pandemic left economic scars of over $10 trillion.[165] Neither of these cases represents superintelligent autonomous adversaries however, and the scope of harms from a true rogue AI may vastly exceed their scale of damage.

A second pathway is slower and more incremental. Instead of a sudden break, control is handed over piece by piece to increasingly capable systems embedded across political, economic, and military systems. Unlike the fast

159 Yoshua Bengio et al., International AI Safety Report 2026, DSIT 2026/001, February 3, 2026, https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026.
160 Ibid.
161 For an example of such a scenario, see: Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean, "AI 2027," April 3, 2025, https://ai-2027.com/.
162 Jan Kulveit et al., "Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development," arXiv, last revised January 29, 2025, https://doi.org/10.48550/arXiv.2501.16946.
163 Alan Chan et al., "Measuring AI R&D Automation," arXiv, last revised March 6, 2026, https://doi.org/10.48550/arXiv.2603.03992.
164 Andy Greenberg, "The Untold Story of NotPetya, the Most Devastating Cyberattack in History," Wired, August 22, 2018, https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/.
165 "Global GDP to Suffer a Loss of $22T Due to COVID-19 Crisis, Says IMF," Forbes Middle East, January 26, 2021, https://www.forbesmiddleeast.com/industry/economy/world-gdp-to-grow-at-55-in-2021-says-imf

takeoff scenario, the risk doesn't come from a single system, but the erosion of meaningful human influence over the systems that structure society. [166]

Distribution of Harm

Across each of these scenarios, it is important to note that consequences are not evenly distributed. Different actors bear different costs, and the effects of AI-enabled failures may be (and are likely to be) felt unevenly across populations and institutions.

Civilian populations are often particularly vulnerable, especially to errors in targeting, misinformation campaigns, and disruptions to critical infrastructure. Errors in targeting, misinformation campaigns, or disruptions to critical infrastructure can produce direct and indirect harms, including loss of life, displacement, and economic disruption. These effects may be exacerbated by the speed and scale of AI-enabled operations.

Democratic accountability represents another dimension of harm. As decision-making becomes more reliant on AI systems, questions arise about responsibility and oversight. When outcomes are shaped by complex interactions between humans and machines, it may be difficult to attribute responsibility for errors or failures. This ambiguity can undermine public trust and complicate governance. Interrelated with this point, there is also a question about how harm may be unevenly distributed based on which countries have the capacity to invest in AI capabilities versus those that may prioritize other domains.

Finally, in the case of military applications of AI technology, alliance dynamics introduce additional complexity—particularly in the context of interoperability as militaries leverage legacy AI systems and integrate new tools.[167] Differences in technological capability, interoperability, and risk tolerance can create friction among partners. Variations in testing standards, data architectures, and governance practices may complicate coordination and reduce the effectiveness of collective action. Failures in one system may have cascading effects across alliance networks, undermining trust and cohesion.

Policymaker Pathologies

The behavior of policymakers under conditions of compressed timelines, opacity, and high stakes represents a critical determinant of consequence. In such environments, decision-makers may exhibit systematic tendencies that amplify risk.

Reliance on technical authority is one such tendency. AI-generated outputs may be treated as more objective or reliable than they are, particularly when decision-makers lack the expertise to evaluate them critically. This can lead to overconfidence in system outputs and insufficient scrutiny of underlying assumptions.

Precautionary logic represents another pathway to escalation. Faced with uncertainty and potential high-consequence risks, decision-makers may adopt a "better safe than sorry" approach, taking actions that are intended to reduce risk but may in fact increase the likelihood of escalation. This logic is particularly powerful in environments where the costs of inaction are perceived to be high.

Reduced tolerance for uncertainty further constrains decision-making. AI systems are often introduced to reduce uncertainty, but their outputs may instead highlight ambiguity or produce conflicting signals. In response, decision-

166 Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud, "Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development," arXiv preprint, arXiv:2501.16946, January 28, 2025, https://arxiv.org/abs/2501.16946

167 Theresa Hitchens, "NATO Needs Policies, Standards for Sharing AI-Enhanced Geospatial Intel: Official," Breaking Defense, May 5, 2026, https://breakingdefense.com/2026/05/nato-needs-policies-standards-for-sharing-ai-enhanced-geospatial-intel-official/ and Sophie Mayo, Impact and Effort: A Menu of AI and Autonomy Options for AUKUS Pillar II (Sydney, Australia: United States Studies Centre, University of Sydney, January 15, 2026), https://www.ussc.edu.au/a-menu-of-ai-and-autonomy-options-for-aukus-pillar-ii

makers may seek to resolve uncertainty through action, rather than through additional analysis or deliberation. Finally, a preference for speed over deliberation may emerge as a dominant organizational norm. As systems enable faster decision-making, the expectation of rapid response may become institutionalized, crowding out slower, more reflective processes.

Given the thin empirical base and the dynamic nature of AI-enabled risks, policymakers will likely require new analytical tools for exploring escalation pathways and testing governance approaches under uncertainty. Wargaming may be particularly useful in this context because many of the risks discussed in this chapter—escalatory misinterpretation, decision compression, and cross-domain entanglement—emerge through interaction, perception, and decision-making under pressure rather than through static technical failure alone.[168]

From Threat to Vulnerability to Consequence

If threat analysis outlines the variation in intent and capabilities, vulnerability analysis identifies where systems are most likely to break, and consequence analysis examines what happens when they do, the central policy question becomes how to maintain meaningful human control, operational visibility, and institutional restraint without sacrificing effectiveness in the context of both commercial and military contexts.

The vanishing margin for error described in this chapter is not an inevitable outcome of AI-enabled military integration. It is the product of choices about how systems are designed, deployed, and governed. Addressing these challenges requires not only technical solutions, but institutional and conceptual ones: rethinking how decisions are made, how risks are managed, and how responsibility is assigned in an environment shaped by increasingly complex human–machine systems.

In this sense, consequence is not merely an endpoint. It is a lens through which to evaluate the adequacy of existing approaches to control, visibility, and restraint—and a guide to where reform is most urgently needed in the pursuit of societal resilience. It is with that in mind that we turn to the policy recommendations over the near- and longer-term that might mitigate risk.

168 Andrew W. Reddie, Bethany L. Goldblum, Kiran Lakkaraju, Jason Reinhardt, Michael Nacht, and Laura Epifanovskaya. "Next-Generation Wargames." Science 362, No. 6421 (2018): 1362-1364 and Erik Lin-Greenberg, Reid BC Pauly, and Jacquelyn G. Schneider. "Wargaming for International Relations Research." European Journal of International Relations 28, No. 1 (2022): 83-109.

Chapter 6. Policy Recommendations

The preceding chapters examined how AI technologies interact with global risk through three components: threat, vulnerability, and consequence (TVC). Threat analysis asks who or what can generate harm, with what intent and capability. Vulnerability analysis examines where systems, institutions, and human-machine interactions are fragile. Consequence analysis considers what happens when those threats pass through those vulnerabilities and produce harm.

In practice, these categories are deeply interconnected. Changes in AI capability can alter the threat landscape while simultaneously increasing vulnerability through insecure deployment or opaque integration. Vulnerabilities in crisis-management systems can also magnify consequences. AI's impact on global risk therefore emerges not from a single variable, but from the interaction among them.

Compounding this complexity, the uncertainty around AI's technological trajectory means that the scale and timing of these risks are uncertain. Policymakers do not know how quickly capabilities will advance, how widely they will diffuse, or which failure modes will dominate. Under these conditions, no single intervention is likely to fully characterize or eliminate risk entirely.

Instead, governments and governance regimes must rely on layered defenses, each targeting different parts of the system or reducing uncertainty about them. Popularized by the "Swiss cheese" model within the organizational safety literature, each layer, from technical safeguards to organizational processes, has gaps.[169] Serious failures occur when those gaps align. Policy recommendations emanating from this analysis, therefore, need to serve two functions: to add and strengthen layers that reduce threat, vulnerability, and consequence, and to improve the evidence base needed to identify where those layers are failing or becoming misaligned.

In this chapter, we lay out such a set of policy recommendations in the form of five pillars built on a foundation of government capacity. Importantly, the value of these pillars differs depending on which view one takes of AI's future trajectory.

If the "mirage" view is substantially correct, policy should prioritize disciplined evidence-gathering, careful procurement, restrictions on inappropriate deployment, and a resistance to hype. Under the "normal technology" view, the central challenge becomes governing diffusion: testing systems before high-stakes deployment, managing human-machine interaction, building institutional capacity, and avoiding brittle integration into sensitive domains. And if AI is instead moving toward increasingly autonomous and transformative systems, many of the same interventions become more urgent while also taking on different strategic significance. Measuring AI R&D automation may function as an early warning system for recursive capability gains, while model governance, security, and resilience planning become tools for managing increasingly powerful systems that may themselves become part of the threat landscape.

The recommendations below are most oriented toward AI systems on an uncertain trajectory between the "normal technology" and "autonomous power" perspectives, where we also find some of the largest potential for global risks.

A summary table of the recommendations and how they impact threat, vulnerability, and consequence can be found at the bottom of the chapter.

The Foundation. Build government capacity, coordination, and translation infrastructure

169 James Reason, Human Error (Cambridge University Press, 1990).

None of the five pillars below can function if the U.S. government lacks the technical capacity and institutional mechanisms needed to understand evolving AI capabilities and respond to emerging threats. Government needs the ability to evaluate frontier systems independently, interpret technical claims made by developers and critics, coordinate across agencies, use AI tools appropriately inside the government, and share risk-relevant information across domestic and international government channels.

For the U.S. government, a stronger AI capacity infrastructure should include at least five components.

- **A SUBSTANTIALLY RESOURCED TECHNICAL EVALUATION AND STANDARDS FUNCTION.** The Center for AI Standards and Innovation (CAISI) in the Department of Commerce should have expanded capacity to conduct rapid, independent AI evaluations for factors such as dangerous capabilities, reliability, and model behavior, as well as advancing the science of AI evaluations and AI standards. These capabilities might come from scaled-up "CAISI+" within Commerce,[170] and a federally funded research and development center supporting the AI mission.[171] This recommendation aligns with the Trump administration's 2025 AI Action Plan,[172] as well as Congressional efforts to put CAISI on statutory footing.[173]
- **A DURABLE TALENT PIPELINE AND SURGE CAPACITY.** Agencies need permanent technical staff who can manage AI integration across their workstreams and engage credibly with frontier AI developers, but they also need a mechanism for rapidly bringing in external AI security expertise during crises.[174] A National AI Reserve Corps, pre-cleared experts, and prearranged contracting mechanisms would help close the gap between crisis timelines and ordinary federal hiring processes.[175]
- **OPERATIONAL TRANSLATION BETWEEN AI DEVELOPERS AND AGENCIES.** Frontier capabilities are concentrated in private firms, while crisis response and public accountability sit largely with government. Government needs trusted mechanisms for model access, evaluation partnerships, and incident reporting that protect sensitive commercial and national-security information while giving agencies enough visibility to act.[176] Past public-private partnership between Anthropic and the National Nuclear Security Administration on nuclear safeguards in AI models illustrates a positive example of such partnerships.[177]
- **GOVERNMENT-TO-GOVERNMENT INFORMATION SHARING ON AI RISKS.** The United States should strengthen channels for sharing AI-risk indicators, evaluation results, incident patterns, and crisis-relevant warnings with other governments. These channels should include both classified and unclassified tracks, common taxonomies for AI incidents and dangerous capabilities, technical expertise and infrastructure for TEVV, and procedures for rapid escalation when AI systems are implicated in risk areas such as cyber activity, concerning biological design workflows, nuclear-related decision support, or military escalation. Information

170 Joe O'Brien, A National Center for Advanced AI Reliability and Security, Federation of American Scientists, June 2025. https://fas.org/publication/a-national-center-for-advanced-ai-reliability-and-security/.
171 David W. Jacobs and Oliver Stephenson, A National AI Laboratory to Support the Administration's AI Agenda at the Department of Commerce, Federation of American Scientists, February 2026. https://fas.org/publication/national-ai-laboratory-at-commerce/.
172 The White House, America's AI Action Plan (Washington, DC: The White House, July 2025), https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf
173 U.S. Senate Committee on Commerce, Science, and Transportation, "Cantwell, Young, Hickenlooper, and Blackburn Reintroduce Bill to Ensure U.S. Leads Global AI Innovation," press release, February 2026. https://www.commerce.senate.gov/press/dem/release/cantwell-young-hickenlooper-and-blackburn-reintroduce-bill-to-ensure-u-s-leads-global-ai-innovation/
174 Enlli Lewis, From Strategy to Impact: Establishing an AI Corps to Accelerate HHS Transformation, Federation of American Scientists, December 2024, https://fas.org/publication/ai-corps-hhs-transformation/.
175 Cara Labrador et al., Building AI Surge Capacity: Mobilizing Technical Talent into Government for AI-Related National Security Crises, Institute for AI Policy and Strategy, October 2025. https://www.iaps.ai/research/building-ai-surge-capacity. See also the Tech Force initiative by the current administration: U.S. Office of Personnel Management, "OPM Launches US Tech Force to Implement President Trump's Vision for Technology Leadership," news release, December 15, 2025, https://www.opm.gov/news/news-releases/opm-launches-us-tech-force-to-implement-president-trumps-vision-for-technology-leadership/.
176 John Croxton et al., Message Incoming: Establish an AI Incident Reporting System, Federation of American Scientists, June 2024, https://fas.org/publication/establishing-an-ai-incident-reporting-system/.
177 Anthropic, Developing Nuclear Safeguards for AI through Public-Private Partnership, August 2025, https://www.anthropic.com/news/developing-nuclear-safeguards-for-ai-through-public-private-partnership.

sharing and collaboration between national AI safety institutes, or equivalent bodies, is a particularly promising area for bolstering short-term collaboration.[178]

- **CLEAR STANDARDS FOR RESPONSIBLE GOVERNMENT USE OF AI.** Government employees should have access to AI tools to gain a clear understanding of their strengths and weaknesses, and use the tools to improve analysis, service delivery, or resilience. However, agencies should not be pushed into using AI where it is inappropriate. Federal policy should define use cases, red lines, verification requirements, and human accountability standards, especially for surveillance, law enforcement, military decision support, and other high-stakes contexts. These policies might then serve as a basis for use cases outside of public service delivery.

Mapped to the TVC framework above, these strengthened government capacities reduce uncertainty and increase risk management across all three terms. It improves threat assessment by helping agencies understand emerging actor capabilities and intent, while also providing situational awareness about increasingly autonomous AI systems. It reduces vulnerability by giving government the technical competence to oversee systems it buys, regulates, or relies on. It reduces consequence by strengthening crisis preparation and response. Without this foundation, the remaining pillars will not be implementable policy levers.

Pillar 1. Build Testing, Evaluation, Verification, and Validation (TEVV) capacity and early-warning systems for dangerous AI capabilities

The first pillar builds the evidence base needed to govern AI under uncertainty. Policymakers cannot avoid making forecasts about AI, but they can make those forecasts more explicit, more measurable, and easier to revise. A national measurement and early-warning system should track not only benchmark performance, but also reliability, autonomy, AI R&D automation, dangerous dual-use capabilities, signs of misalignment, scale of deployment, components of the AI supply chain, and the conditions under which model capabilities translate into real-world risk.

The base of this pillar is testing, evaluation, verification, and validation (TEVV). Testing is the empirical process of observing system behavior under specified conditions. Evaluation is the broader assessment of a system against defined criteria. Verification asks whether the system was built correctly according to requirements ("did you build the system right?"). Validation asks whether it is fit for the intended real-world purpose ("did you build the right system?"). For AI technologies, all four remain underdeveloped. Benchmark performance is often treated as a proxy for general capability, but many benchmarks have weak construct validity: they measure something precisely without necessarily measuring the property that matters for policy. NIST's proposed TEVV standard is an important starting point, but the science of AI evaluation needs much more sustained investment.[179]

A TEVV and early-warning agenda should include several categories of work:

- **TRAJECTORY INDICATORS.** Government should support monitoring and measurements[180] that help distinguish among the trajectories described in Chapter 1: hype and overdeployment, diffusion of AI in a

178 For an example of incident reporting legislation, see: U.S. Congress, House, AI Incident Reporting and Security Enhancement Act, H.R. 9720, 118th Cong., 2nd sess., introduced September 20, 2024, https://www.govinfo.gov/app/details/BILLS-118hr9720ih. On confidence building measures and international cooperation see: Michael Horowitz and Paul Scharre, AI and International Stability: Risks and Confidence-Building Measures (Washington, DC: Center for a New American Security, January 12, 2021), https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures; Richard Danzig, Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority (Washington, DC: Center for a New American Security, May 30, 2018), https://www.cnas.org/publications/reports/technology-roulette.

179 National Institute of Standards and Technology (NIST), Outline: Proposed Zero Draft for a Standard on AI Testing, Evaluation, Verification, and Validation (TEVV), July 2025, https://www.nist.gov/system/files/documents/2025/07/15/Outline_%20Proposed%20Zero%20Draft%20for%20a%20Standard%20on%20AI%20TEVV-for-web.pdf.

180 Jess Whittlestone and Jack Clark, "Why and How Governments Should Monitor AI Development," arXiv, last revised August 31, 2021, https://doi.org/10.48550/arXiv.2108.12427.

"normal technology" model, and more transformative forms of automation. Relevant indicators worth tracking could include:

  - Trends in key components of the AI supply chain (for example data center construction, AI energy consumption, AI chip production, semiconductor manufacturing equipment, training data availability).
  - Rates of AI use in different industries, and their impacts (e.g. consequences for productivity and employment).
  - Rate of AI research and development automation (which is one plausible mechanism for faster AI capability progress and shorter policy windows).[181]

- **CAPABILITY-RELIABILITY MEASUREMENT.** Evaluating AI systems on a small number of numerical benchmarks often misses real-world failure modes. Evaluation should ask not only whether a model can do a task once, but whether a deployed system can do it reliably, under distribution shift, adversarial pressure, and realistic time and resource constraints.[182] These evaluations should increasingly consider the performance of human-AI systems, rather than just the AI systems themselves.
- **THREAT-MODEL-SPECIFIC EVALUATIONS.** AI should be evaluated against concrete threat pathways, for example the Cyber Kill Chain,[183] or biosecurity workflows. These evaluations should consider actor access, intent, and capability together. They should also incorporate government's best estimates of where the genuine bottlenecks in a given process are and to what extent AI reduces those bottlenecks.[184]
- **EVALUATION-TO-RISK TRANSLATION.** Policymakers need methods for converting evaluation results into estimates of risk.[185] Such work should draw on threat-model-specific evaluations outlined above. The goal is to go beyond model performance on a particular benchmark and connect the amount of uplift a model can provide in particular areas to changes in real-world risk.
- **POST-DEPLOYMENT MONITORING AND INCIDENT LEARNING.** AI capabilities are often difficult to fully characterize before widespread deployment, meaning pre-deployment evaluations alone are insufficient.[186] This gap can be addressed by a combination of developing voluntary standards for first-party post-deployment monitoring, and building government's ability to do such monitoring directly.
- **POLICY-RELEVANT CAPABILITY MILESTONES AND THRESHOLDS FOR POWERFUL AI.** Terms such as "AGI" or "ASI" are often used as important thresholds for transformative change, but they have a range of competing definitions—some of which are tied up in business outcomes.[187] Policymakers should direct the creation of more granular thresholds for particularly transformative capabilities. Abilities such as long-horizon

181 Alan Chan et al., Measuring AI R&D Automation, GovAI, March 2026. https://www.governance.ai/research-paper/measuring-ai-r-d-automation.
182 Stephan Rabanser et al., Towards a Science of AI Agent Reliability, arXiv, February 18, 2026, https://doi.org/10.48550/arXiv.2602.16666.
183 Lockheed Martin, Cyber Kill Chain®, Lockheed Martin, accessed April 26, 2026, https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html.
184 For examples of a detailed analysis of bottlenecks, see: Sonia Ben Ouagrham-Gormley, Barriers to Bioweapons: The Challenges of Expertise and Organization for Weapons Development (Ithaca, NY: Cornell University Press, 2014), https://www.cornellpress.cornell.edu/book/9780801452888/barriers-to-bioweapons/. For a more concrete risk model from powerful cyber capabilities see: John Halstead and Luca Righetti, Assessing the Risk of AI-Enabled Computer Worms, GovAI, September 2025. https://www.governance.ai/research-paper/assessing-the-risk-of-ai-enabled-computer-worms.
185 Luca Righetti, Dual-Use AI Capabilities and the Risk of Bioterrorism: Converting Capability Evaluations to Risk Assessments, GovAI, December 2025. https://www.governance.ai/research-paper/dual-use-ai-capabilities-and-the-risk-of-bioterrorism-converting-capability-evaluations-to-risk-assessments.
186 Anita Rao et al., Challenges to the Monitoring of Deployed AI Systems: Center for AI Standards and Innovation, NIST AI 800-4, March 2026. https://www.nist.gov/publications/challenges-monitoring-deployed-ai-systems-center-ai-standards-and-innovation.
187 Dan Hendrycks et al., A Definition of AGI, arXiv, 2025. https://arxiv.org/abs/2510.18212; Helen Toner, The Term 'AGI' Is Almost Useless at This Point, Rising Tide (Substack), April 6, 2026. https://helentoner.substack.com/p/the-term-agi-is-almost-useless-at; Meredith Ringel Morris et al., "Levels of AGI for Operationalizing Progress on the Path to AGI," arXiv, last revised September 24, 2025, https://doi.org/10.48550/arXiv.2311.02462.

autonomy,[188] sophisticated tool use, and self-sufficiency[189] could be relevant thresholds. Progress towards such thresholds could then be included in the trajectory indicators described above.

- **MODEL BEHAVIOR AND MISALIGNMENT MEASUREMENT.** Measurement of AI systems should increasingly include phenomena such as scheming, sandbagging, evaluation awareness, self-exfiltration attempts, deception, and resistance to shutdown or modification. Pre-deployment evaluations become less reliable if models can alter their behavior when they are being tested, and such behaviors could be early signs of more dangerous misalignment.[190] At the same time, the approaches for making such measurements are still being developed, and results require careful interpretation.[191]

This pillar maps to our TVC analysis above in two ways. First, it reduces uncertainty across threat, vulnerability, and consequence. Better measurement helps identify emerging actor affordances, dangerous capabilities, and possible warning signs. Second, it gives defenders lead time. If we can detect movement toward more transformative trajectories—for example, rapid AI R&D automation or improved autonomous cyber capabilities—we can adjust the prioritization between the other policy pillars before the window for action closes. However, a measurement system risks becoming "dashboard theater" if indicators are not converted into actionable policy levers, such as those laid out below.

Pillar 2. Govern the technical layer: compute, models, weights, access, safeguards, and dangerous capabilities

The second pillar focuses on the technical layer of AI systems: compute, model behavior, model weights, access controls, safeguards, training data, monitoring, and dangerous capability thresholds. Technical-layer governance is not sufficient on its own, because AI safety is partly a property of deployment context and human use, but it remains essential. Earlier intervention at the technical layer can prevent dangerous capabilities, insecure deployments, or exploitable vulnerabilities from propagating downstream into settings where they are harder to control.

The technical layer includes at least five kinds of policy interventions:

- **COMPUTE GOVERNANCE.** Compute is a foundational input to AI training and deployment, and restrictions on advanced AI chips can limit the capabilities and diffusion of frontier AI models.[192]
- **MODEL WEIGHT SECURITY.** Advanced model weights can be high-consequence assets. If stolen or openly released without adequate safeguards, they may allow actors to remove safety filters, fine-tune models for harmful uses, or reproduce capabilities that would otherwise remain limited. Model-weight protection should therefore be treated as part of national-security risk management for frontier systems, with requirements that scale with capability and misuse potential.[193]

188 Thomas Kwa et al., Measuring AI Ability to Complete Long Tasks, METR Blog, March 19, 2025, https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/.
189 Ajeya Cotra, "Self-Sufficient AI," Planned Obsolescence (Substack), January 6, 2026, https://www.planned-obsolescence.org/p/self-sufficient-ai.
190 Sambhav Maheshwari and Joe O'Brien, Evaluation Awareness: Why Frontier AI Models Are Getting Harder to Test, Institute for AI Policy and Strategy, March 2026, https://www.iaps.ai/research/evaluation-awareness-why-frontier-ai-models-are-getting-harder-to-test; Anthropic, Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training, January 2024, https://www.anthropic.com/news/sleeper-agents-training-deceptive-llms-that-persist-through-safety-training.
191 Christopher Summerfield et al., Lessons from a Chimp: AI "Scheming" and the Quest for Ape Language, UK AI Security Institute, 2025, https://www.aisi.gov.uk/research/lessons-from-a-chimp-ai-scheming-and-the-quest-for-ape-language.
192 Girish Sastry et al., "Computing Power and the Governance of Artificial Intelligence," arXiv preprint arXiv:2402.08797, February 13, 2024, https://arxiv.org/abs/2402.08797.
193 Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott, Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models (Santa Monica, CA: RAND Corporation, 2024), https://doi.org/10.7249/RRA2849-1.

- **DANGEROUS-CAPABILITY CONTROLS.** Policymakers should support methods for reducing or restricting potentially dangerous capabilities, particularly where they are not needed for beneficial use. This includes research on safety fine-tuning, output filtering, model access restrictions, and techniques such as unlearning.[194]
- **MONITORABILITY AND INTERPRETABILITY.** For increasingly agentic systems, policymakers should support research into interpretability, robustness, control, and the preservation of monitorable AI reasoning where it materially improves oversight.[195]
- **SAFETY PLANS AND TRANSPARENCY OBLIGATIONS.** Frontier AI developers should have written safety and security frameworks that explain how they identify dangerous capabilities, what thresholds trigger additional safeguards, and how they handle serious incidents.[196] California's SB 53 and New York's RAISE Act provide state-level examples of transparency, safety framing, and incident-reporting approaches. They should be treated as evidence-generating governance mechanisms, not as final federal models. [197],[198]

Technical-layer governance maps most directly to threat and vulnerability. It can reduce threat by limiting the capabilities available to malicious actors. It can reduce vulnerability by decreasing the likelihood that insecure model behavior, compromised weights, or poorly understood failure modes enter high-stakes systems. It can reduce consequence indirectly by preventing some high-consequence misuse pathways from becoming available in the first place. The main caveat is that technical controls can be brittle. For example, they can be bypassed or eroded by diffusion of open-weight models without controls. For that reason, this pillar must be paired with other components of this plan.

Pillar 3. Govern deployment in sociotechnical systems, not just models

The third pillar focuses on the environments in which AI systems are deployed. In today's systems, the relevant unit of analysis is usually not the model alone, but the human-AI system: the model, the interface, the user, the organization, the workflow, the incentives, and the decision processes surrounding it.

Sociotechnical deployment governance should be risk-tiered. Low-risk administrative use cases should not face the same requirements as systems that materially influence kinetic decisions, biological design workflows, cyber operations, or crisis response. But high-risk deployments should face more stringent requirements: independent evaluation, continuous monitoring, audit logs, human-factors testing, clear lines of accountability, tested fallback procedures, and explicit constraints on system behavior.

Priorities here include:

- **PRESERVE MEANINGFUL HUMAN CONTROL, NOT MERELY FORMAL HUMAN-IN-THE-LOOP REVIEW.** A human can be in the loop and still function as a rubber stamp. Deployment rules should assess whether

194 Nathaniel Li et al., "The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning," Proceedings of Machine Learning Research 235, 2024, https://proceedings.mlr.press/v235/li24bc.html

195 Oscar Delaney, Oliver Guest, and Renan Araujo, Policy Options for Preserving Chain of Thought Monitorability, Institute for AI Policy and Strategy, September 2025, https://www.iaps.ai/research/policy-options-for-preserving-cot-monitorability; Matteo Pistillo, "Accelerating AI Interpretability to Promote U.S. Technological Leadership," Federation of American Scientists, June 10, 2025, https://fas.org/publication/accelerating-ai-interpretability/.

196 E.g. see existing plans by leading AI companies: Anthropic, Responsible Scaling Policy, version 3.2, effective April 29, 2026, https://cdn.sanity.io/files/4zrzovbb/website/28c6241900d90410628a8a2003a5572faae4365a.pdf; Google DeepMind, Frontier Safety Framework, version 3.1, published April 17, 2026, https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3-1.pdf; OpenAI, Preparedness Framework, version 2, last updated April 15, 2025, https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf.

197 California Governor's Office, Governor Newsom signs SB 53, advancing California's world-leading artificial intelligence industry, September 2025, https://www.gov.ca.gov/2025/09/29/governor-newsom-signs-sb-53-advancing-californias-world-leading-artificial-intelligence-industry/

198 New York State Department of Financial Services, Governor Hochul Signs Nation-Leading Legislation to Require AI Frameworks for AI Frontier Models, December 2025, https://www.dfs.ny.gov/reports_and_publications/press_releases/pr20251222

human operators have the time, information, authority, and training needed to understand system limits and override outputs.[199]

- **MANAGE BOTH AUTOMATION BIAS AND ALGORITHMIC AVERSION.** Operators may over-rely on AI outputs when systems appear objective, fast, or authoritative; they may also under-rely on useful systems after visible failures. Both failure modes matter. Training should build calibrated trust by teaching users the capability-reliability gap and the conditions under which systems fail.
- **REDUCE TIGHT COUPLING AND HIDDEN COMPLEXITY WHERE POSSIBLE.** Drawing on normal accident theory, AI should not be integrated into complex, tightly coupled workflows without strong justification. Where tight coupling is unavoidable, the burden on testing, redundancy, monitoring, and human intervention should increase.
- **CREATE A ROBUST AI INCIDENT REPORTING SYSTEM.** A national AI incident reporting hub could help government and industry learn from failures, near misses, and dangerous capability discoveries.[200]
- **USE WARGAMING AND EXERCISES FOR MAPPING DYNAMIC RISKS.** Many AI-global risks arise from interaction, perception, and decision-making under pressure. Wargaming can help examine escalation, cross-domain entanglement, and organizational response in ways that static model evaluations cannot.

This pillar maps strongly onto vulnerability and consequence. It reduces vulnerability by making deployment environments less brittle: fewer hidden failure modes, less unchecked delegation, better human-machine interaction, and clearer accountability. It reduces consequence by preserving deliberation, visibility, and restraint in settings where errors could otherwise cascade. It can also reduce threat by limiting accidental, delegated, or organizationally generated pathways to harm—cases where no malicious actor is necessary for risk to emerge.

Pillar 4. Shift the offense-defense balance of AI systems toward defense

The fourth pillar is to shape AI development and deployment so that defensive uses mature faster than offensive ones.[201] AI is dual-use: the same capabilities that help defenders identify vulnerabilities, advance research, or respond to threats may also help attackers discover exploits, design harmful agents, or evade safeguards. The policy objective cannot simply be to slow all AI capability. It should be to differentially accelerate protective capabilities, give defenders earlier and safer access, and delay or restrict access to capabilities whose immediate broad release would advantage attackers.

Practical policy responses could include:

- **DEFENSIVE AI FOR CYBER.** Government should fund, test, and diffuse AI tools for vulnerability discovery, patch prioritization, configuration management, alert triage, and secure code review. DARPA-style challenges,

199 For sources, please see U.S. Department of Defense, DoD Directive 3000.09: Autonomy in Weapon Systems (Washington, DC: Department of Defense, January 25, 2023) https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf and U.S. Department of State, "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," Bureau of Arms Control, Deterrence, and Stability, 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. For further discussion of the limitations of "human in the loop" see part V of: Richard Danzig, Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority (Washington, DC: Center for a New American Security, May 30, 2018), https://www.cnas.org/publications/reports/technology-roulette.

200 John Croxton, David Robusto, Satya Thallam, and Doug Calidas, Establish an AI Incident Reporting System, Federation of American Scientists, June 2024, https://fas.org/publication/establishing-an-ai-incident-reporting-system/.

201 Jonas Sandbrink, Hamish Hobbs, Jacob Swett, Allan Dafoe, and Anders Sandberg, "Differential Technology Development: An Innovation Governance Consideration for Navigating Technology Risks," SSRN, September 8, 2022, last revised December 3, 2023, https://doi.org/10.2139/ssrn.4213670.

cyber ranges, government pilots, and procurement incentives can help move defensive AI from demos into operational use.[202]

- **AI-ENABLED BIOSECURITY.** AI can strengthen DNA synthesis screening, anomaly detection in biological supply chains, biosurveillance, and risk analysis at the digital-physical boundary. These defensive tools should be developed and deployed before broadly available AI systems substantially lower barriers to harmful biological workflows.
- **DIFFERENTIAL ACCESS FOR DEFENDERS.** One approach here is to promote low-risk defensive capabilities widely, manage access to medium-risk capabilities, and deny-by-default access to the highest-risk capabilities while preserving access for trusted defenders. [203] Government can support this through evaluation standards, procurement, grants, cyber ranges, and incentives for AI developers to prioritize strategically important defenders.
- **STAGED RELEASES AND DEFENDER-FIRST ACCESS.** It will often not be possible to permanently restrict the highest-risk capabilities given the rapid proliferation of AI capabilities. When models appear likely to improve offensive capabilities but proliferate quickly, developers should consider staged deployment, restricted release, or early access for vetted defenders, critical infrastructure operators, and government evaluators. This will at least give defensive institutions additional time to adapt.

This pillar maps most directly to threat and vulnerability. It reduces threat by raising attacker costs, reducing attacker return on investment, and limiting immediate access to high-risk capabilities. It reduces vulnerability by hardening critical systems before offensive capability diffuses. It reduces consequence by slowing the speed and scale of harm when attacks occur, giving defenders and responders more time. The main tradeoff is that access restrictions can also delay beneficial uses. Policymakers should therefore define clear criteria for capability tiers, defender eligibility, sunset reviews, and mechanisms for expanding access when risks are mitigated.

Pillar 5. Build societal and institutional resilience for when prevention fails

The fifth pillar starts from the premise that some AI-enabled risks will proliferate, some safeguards will fail, and some incidents will occur despite the best prevention efforts. This is especially true if open-weight systems remain close to the frontier, if offensive capabilities diffuse faster than defensive institutions can absorb them, or if increasingly capable and autonomous systems create failure modes that are difficult to anticipate.

Examples of building resilience include:

- **GENERAL CYBER RESILIENCE.** Critical infrastructure, state and local governments, hospitals, utilities, and election systems need stronger baseline cybersecurity because AI-enabled attackers may exploit known weaknesses faster and at lower cost. Basic hardening, incident response, backup systems, segmentation, and recovery planning are no-regrets investments. Government agencies like the Cybersecurity and Infrastructure Security Agency have a key role to play here.
- **BIOSAFETY AND BIOSECURITY RESILIENCE.** DNA synthesis screening, bio lab safety, public health surveillance, medical countermeasure platforms, and outbreak response all reduce the consequence of AI-enabled biological misuse or accidents. These measures matter even if the most extreme AI-bio risks do not materialize.

202 For example, see: Defense Advanced Research Projects Agency, "AI Cyber Challenge (AIxCC)," accessed May 8, 2026, https://aicyberchallenge.com/; Cybersecurity and Infrastructure Security Agency, "Joint Cyber Defense Collaborative," accessed May 11, 2026, https://www.cisa.gov/topics/partnerships-and-collaboration/joint-cyber-defense-collaborative.

203 Shaun Ee et al., Asymmetry by Design: Boosting Cyber Defenders with Differential Access to AI, Institute for AI Policy and Strategy, May 2025; Christopher Covino and Shaun Ee, Policy Actions for Enabling Cyber Defense Through Differential Access, August 2025, https://www.iaps.ai/research/policy-actions-for-enabling-cyber-defense-through-differential-access

- **NUCLEAR AND MILITARY RISK REDUCTION.** There are salient concerns that the integration of AI technologies into military contexts might compress decision-making timelines, erode meaningful human control, or create unjustified confidence in early warning and command systems. Efforts need to be undertaken to ensure that AI systems do not erode crisis stability or strategic stability. Resilience in this context includes guardrails concerning use in military contexts, crisis communication, escalation management, and confidence-building measures with both allies and competitors.[204]
- **CRISIS COMMUNICATION AND PUBLIC TRUST.** AI-enabled incidents may create confusion, false attribution, panic, or competing narratives. Governments need pre-planned communication protocols, evidence standards for attribution, and coordination mechanisms across agencies and international partners.
- **WORST-CASE SCENARIO PLANNING.** Government should exercise scenarios involving AI-enabled cyber worms, AI-assisted WMD pathways, AI-driven military escalation, and potential loss-of-control incidents. These exercises should identify decision points, legal authorities, emergency contacts, and gaps in technical response.
- **CONTAINMENT AND INTERRUPTION MECHANISMS.** For increasingly autonomous AI systems, policymakers should support research into control, containment, logging, emergency shutdown, and rollback mechanisms. These control mechanisms should be embedded in consequential AI deployments, and governments should plan for how to use them in the event of a loss-of-control scenario.

Resilience maps most directly to consequence. It reduces casualties, cascading failures, panic, and escalation. Such interventions may also reduce vulnerability. Its effect on threat is less direct: if attackers expect lower payoff because societies recover quickly and can quickly attribute harms, some attacks become less attractive.

Resilience should not be used as an excuse to tolerate preventable hazards upstream. The five-pillar structure is designed to avoid that mistake: resilience complements measurement, technical governance, deployment governance, and defensive advantage. It does not replace them.

Putting the pillars together

The five pillars should not be read as a sequence in which government completes one step before beginning the next. They are mutually dependent. Measurement informs model governance, deployment rules, defensive access, and resilience planning. Technical controls reduce some risks before deployment, but sociotechnical governance determines whether those controls survive contact with real institutions. Defensive advantage and resilience matter because some capabilities will diffuse and some prevention will fail. Government capacity is the enabling layer that makes the other pillars operational.

204 Michael Horowitz and Paul Scharre, AI and International Stability: Risks and Confidence-Building Measures (Washington, DC: Center for a New American Security, January 12, 2021), https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures; Richard Danzig, Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority (Washington, DC: Center for a New American Security, May 30, 2018), https://www.cnas.org/publications/reports/technology-roulette.

Summary. Mapping the Five Pillars and Foundation onto Threat, Vulnerability, and Consequence

The following pages summarize how each recommendation layer contributes to the risk management across threat, vulnerability, and consequence, as well as reducing uncertainty across all three. They also briefly describe key policy implementation challenges.

Foundation. Government capacity, coordination, and translation infrastructure

CORE POLICY OBJECTIVE AND EXAMPLES

Build the capacity that makes the other pillars operational: a stronger CAISI/standards and evaluation function; durable AI talent and surge capacity; public-private model access, evaluation partnerships, and incident reporting; government-to-government risk sharing; and clear standards for responsible government use of AI.

THREAT

Improves threat assessment by giving agencies better visibility into actor capabilities and intent, malicious-use patterns, dangerous capabilities, and increasingly autonomous AI systems.

VULNERABILITY

Reduces institutional fragility by giving government the technical competence to oversee systems it buys, regulates, or relies on; reduces fragmented ownership and weak technical review.

CONSEQUENCE

Strengthens crisis preparation, incident response, diplomatic coordination, and responsible government use in high-stakes contexts.

UNCERTAINTY REDUCTION

Creates the information channels and independent technical capacity needed to reduce uncertainty across all three TVC terms: what capabilities are emerging, who can use them, and where they may propagate.

KEY IMPLEMENTATION QUESTIONS

- Where should capacity sit: CAISI+, a Commerce FFRDC, an NSC process, DHS/CISA, DoD, DOE, the intelligence community, or a hybrid?
- What information can be shared with allies while protecting sources, methods, and proprietary data?
- What authority and resources are needed?

Pillar 1. Measurement, TEVV, and early warning

CORE POLICY OBJECTIVE AND EXAMPLES

Build the evidence base for governing AI under uncertainty: construct-valid TEVV; trajectory indicators; capability-reliability measurement; threat-model-specific evaluations; evaluation-to-risk translation; post-deployment monitoring; policy-relevant thresholds for powerful AI; and measurement of model behavior or misalignment.

THREAT

Detects dangerous capabilities and actor affordances before broad misuse; identifies when human, human-AI, or AI-as-actor threat models are becoming more plausible.

VULNERABILITY

Gives defenders lead time to strengthen weak layers; identifies brittle deployments, unreliable systems, poorly measured risks, and places where model capabilities translate into real-world failure modes.

CONSEQUENCE

Avoids strategic surprise and allows other policy tools to activate earlier, before harms cascade or policy windows close.

UNCERTAINTY REDUCTION

This is the central uncertainty-reduction pillar. It turns implicit forecasts into measurable indicators and reduces uncertainty about capability trajectories, reliability, thresholds, diffusion, and real-world risk translation.

KEY IMPLEMENTATION QUESTIONS

- Which indicators should trigger policy action?
- Who validates evaluation results?
- How are sensitive results protected and shared?
- How do policymakers avoid benchmark theater and dashboard theater?

Pillar 2. Technical-layer governance

CORE POLICY OBJECTIVE AND EXAMPLES

Govern compute, models, weights, access, safeguards, dangerous capabilities, monitorability, interpretability, developer safety frameworks, transparency obligations, and serious-incident processes before risks propagate downstream.

THREAT

Limits adversary access to high-risk capabilities; reduces model theft, removal of safeguards, and fine-tuning or reuse of models for harmful purposes.

VULNERABILITY

Reduces insecure model behavior, compromised weights, brittle safeguards, and untested dangerous capabilities entering high-stakes systems.

CONSEQUENCE

Prevents some high-consequence misuse pathways from becoming available in the first place; lowers the chance that technical failures propagate into downstream systems.

UNCERTAINTY REDUCTION

Improves knowledge of model behavior, capability thresholds, incidents, and failure modes. The uncertainty reduction is conditional: controls are brittle if diffusion, open-weight release, or bypass methods erode them.

KEY IMPLEMENTATION QUESTIONS

- What capabilities justify heightened controls?
- How should open-weight benefits be weighed against misuse risk?
- What transparency should be public versus confidential to government?
- How can controls be updated as capabilities diffuse?

Pillar 3. Sociotechnical deployment governance

CORE POLICY OBJECTIVE AND EXAMPLES

Govern how AI systems are embedded in institutions: risk-tiered deployment; independent review for high-risk uses; continuous monitoring; audit logs; human-factors testing; fallback procedures; meaningful human control; incident reporting; and wargaming or exercises for dynamic risks.

THREAT

Reduces accidental, delegated, or organizationally generated threat pathways; limits harmful use cases even when underlying models remain available.

VULNERABILITY

Prevents automation bias, algorithmic aversion, function creep, brittle integration, tight coupling, hidden complexity, and unclear accountability.

CONSEQUENCE

Preserves deliberation, visibility, and restraint; reduces escalation, irreversible decisions, loss-of-control incidents, and institutional overreaction under compressed timelines.

UNCERTAINTY REDUCTION

Post-deployment monitoring, incident reporting, exercises, and wargaming reveal real-world human-AI interaction effects that static model evaluations can miss.

KEY IMPLEMENTATION QUESTIONS

- What constitutes meaningful human control in each domain?
- Which deployments require independent review?
- Who is accountable for model updates and downstream failures?
- How can rules distinguish low-risk administrative use from high-risk use in military, bio, cyber, or crisis settings?

Pillar 4. Defensive advantage

CORE POLICY OBJECTIVE AND EXAMPLES

Shape development and deployment so defensive uses mature faster than offensive ones: fund and diffuse defensive AI for cyber; strengthen AI-enabled biosecurity; use differential access; stage releases; provide defender-first access; and define capability tiers and sunset reviews.

THREAT

Raises attacker costs, reduces attacker return on investment, and delays broad access to capabilities whose immediate release would advantage offense.

VULNERABILITY

Hardens critical systems before offensive capability diffuses; gives trusted defenders earlier access to tools for vulnerability discovery, patching, biosurveillance, synthesis screening, and risk analysis.

CONSEQUENCE

Slows the speed, scale, and persistence of harm; gives responders more time to detect, attribute, patch, recover, and communicate.

UNCERTAINTY REDUCTION

Defender-first access, cyber ranges, pilots, exercises, and sunset reviews generate evidence about defensive performance, access tiers, and the timing of capability diffusion.

KEY IMPLEMENTATION QUESTIONS

- Who qualifies as a trusted defender?
- How are access tiers reviewed?
- When should restrictions expire?
- How can defender-first access avoid favoritism or capture while preserving beneficial uses?

Pillar 5. Resilience

CORE POLICY OBJECTIVE AND EXAMPLES

Prepare for prevention failure: general cyber resilience; biosafety and biosecurity resilience; nuclear and military risk reduction; crisis communication and public trust; worst-case scenario planning; and containment, interruption, logging, shutdown, and rollback mechanisms for consequential autonomous systems.

THREAT

Less direct, but real: if recovery, attribution, and containment reduce expected payoff, some attacks become less attractive.

VULNERABILITY

Improves redundancy, response systems, recovery capacity, containment mechanisms, and institutional preparedness across technical and societal layers.

CONSEQUENCE

Most directly reduces casualties, cascading failures, panic, escalation, and long-term institutional damage when prevention fails.

UNCERTAINTY REDUCTION

Exercises, incident learning, near-miss analysis, communication protocols, and containment planning clarify failure modes, legal authorities, operational contacts, and assumptions about loss-of-control scenarios.

KEY IMPLEMENTATION QUESTIONS

- How do policymakers avoid using resilience as an excuse to tolerate preventable upstream risks?
- Which worst-case scenarios deserve exercises now?
- What legal authorities and emergency contacts are needed?
- How should containment mechanisms be embedded in consequential deployments?

Conclusion

Artificial intelligence is not the first technology to reshape global risk. However, it is the first to do so across domains, at speed, and within systems that are already complex, tightly coupled, and difficult to govern.

For much of the modern era, global risk has been approached through relatively bounded domains. Nuclear risks were governed through arms control, deterrence, and verification regimes. Biological risks were managed through public health systems, biosecurity norms, and international coordination. Cyber risk, while more diffuse, still generally preserved some separation between systems, actors, and response mechanisms.

AI does not fit cleanly into any of these models. It is a general-purpose capability that interacts with existing risks rather than replacing them. In doing so, it introduces dynamics that are harder to isolate, slower to understand, and faster to propagate.

The most important shift is from domain-specific risk to system-level risk. In a system-level risk environment, failures are less likely to be discrete and attributable. They are more likely to emerge from interactions across systems, institutions, and actors. A vulnerability in one domain can be amplified by capabilities in another. A system that performs well in isolation may behave unpredictably when integrated into a broader workflow. A decision made under time pressure may propagate across systems before it can be meaningfully assessed or reversed.

We see these issues emerging in the gap between what AI systems can demonstrate in controlled environments and what they can reliably do in real-world settings. We see them in the "jagged" nature of capability, where systems can perform at a high level in one domain while failing in another that appears similar. We see them in the growing difficulty of evaluating systems whose behavior changes depending on context, interaction, and deployment environment.

Across government, industry, and research environments, AI is being integrated into workflows that were not designed for it. In some cases, this integration is thoughtful and measured. In others, it is driven by competitive pressure, perceived necessity, or simple inertia. In both cases, the result is the same: AI is becoming embedded in the systems that underpin economic activity, national security, and public decision-making.

Capability is advancing faster than institutions can evaluate, govern, and absorb it. This matters because policymakers are being asked to make decisions under conditions of persistent uncertainty, often before reliable evidence about AI's real-world impacts has emerged.

This is the "evidence dilemma" described in the report: act too early and risk getting it wrong, or act too late and risk being unable to correct course.

There is no way to eliminate this dilemma. But there are ways to manage it.

What This Means for Policymakers

One of the core messages of this report is that policymaking in this space cannot depend on a single view of AI's future.

The three perspectives outlined earlier—AI as overhyped, AI as a normal technology, and AI on a path to powerful autonomous systems—are not going to converge into a single consensus in the near term. Each perspective draws on different evidence, different analogies, and different assumptions about how technology evolves.

Policymakers do not have the luxury of waiting for that convergence. Instead, they must focus on building approaches that remain useful across multiple plausible futures, understanding which future we are moving towards, and being ready to respond.

That means a few things in practice.

First, it means taking uncertainty seriously, but not allowing it to become a reason for inaction. Every policy choice already embeds assumptions about what AI is and where it is going. The task is to make those assumptions explicit, test them against evidence, and revise them as conditions change.

Second, it means focusing on what can be shaped, rather than what can be predicted.

The threat–vulnerability–consequence framework is useful here, not because it provides a precise model of risk, but because it highlights points of intervention. Policymakers may not be able to control how quickly AI capabilities advance. But they can influence how those capabilities are deployed, where vulnerabilities persist, and how consequences are mitigated when failures occur.

Third, it means recognizing that institutions matter as well as technology.

Many of the risks described in this report do not arise from AI systems in isolation. They arise from how those systems are embedded in real-world contexts: how they are used, how they are trusted, how they are overseen, and how they interact with human decision-making.

Strengthening those institutional contexts—through better evaluation, clearer accountability, stronger oversight, and improved coordination—is not as visible as advancing frontier capabilities. But it is often where the most meaningful risk reduction occurs.

Where Risk Reduction Is Most Tractable

Across the report, several areas emerge where policymakers have meaningful leverage.

GOVERNMENT CAPACITY IS FOUNDATIONAL. The ability to understand, use, and oversee AI systems cannot be outsourced entirely to the private sector. Agencies need the technical expertise and institutional mechanisms to engage directly with these systems, interpret claims about their capabilities, and coordinate responses across domains.

MEASUREMENT AND EVALUATION IS AN IMMEDIATE TASK. Without better ways to assess AI systems, policymakers are operating with incomplete and potentially misleading information. Investments in testing, evaluation, verification, and validation are not simply technical exercises. They are foundational to governance.

TECHNICAL GOVERNANCE IS A KEY FIRST LINE OF DEFENSE. Factors including compute, model weights, dangerous capabilities, and model interpretability represent areas where interventions can manage risks before they propagate downstream.

DEPLOYMENT GOVERNANCE IS ALSO CRITICAL. Many of the highest-consequence risks will not come from the existence of AI systems, but from how they are used in high-stakes environments: military decision support, cyber operations, biological research, and critical infrastructure. Ensuring that deployment in these contexts is deliberate, evaluated, and accountable is a central policy task.

THERE IS A PRESSING NEED TO SHIFT THE BALANCE BETWEEN OFFENSE AND DEFENSE. AI can lower the cost of harmful activity in areas like cyber operations and information manipulation. It can also strengthen defensive capabilities. Policy should focus on ensuring that defensive applications keep pace with, and ideally outstrip, offensive uses.

Finally, resilience is often underemphasized, but essential. Even well-governed systems will fail. The ability to detect, respond to, and recover from those failures will shape outcomes as much as the ability to prevent them.

AI does not eliminate the need for human judgment, institutional restraint, or international coordination. If anything, it increases their importance. The challenge facing policymakers is not simply knowing whether AI systems

become more capable, but whether societies can adapt their institutions quickly enough to govern technologies that increasingly shape perception, decision-making, and power across domains. That challenge will not be solved through a single policy intervention or technical safeguard. It will require sustained investment in evaluation, resilience, governance capacity, and international coordination under conditions of persistent uncertainty. The future of AI and global risk will ultimately depend not only on what these systems can do, but on the choices humans make about how they are developed, deployed, and constrained.

Appendix A. Summary of Project

Origins and Purpose

The AIxGlobal Risk (AIxGR) project was launched by the Federation of American Scientists (FAS), in partnership with the Future of Life Institute (FLI), to address a growing challenge at the intersection of artificial intelligence and global security: while AI capabilities were advancing rapidly across multiple domains, the policy and expert communities responsible for understanding and governing those risks often remain fragmented.

Discussions about AI and global risk were frequently siloed by domain. Nuclear experts focused on strategic stability and command-and-control systems. Biosecurity experts narrowly examined how AI could shape biological weapons design and information hazards. Cybersecurity practitioners focused on automation, scale, and persistent exploitation. AI researchers debated the pace of capability development, evaluation, and alignment. And policymakers and civil society organizations were often tasked with navigating these issues simultaneously, despite the absence of both a shared framework and common language across these communities.

The AIxGR initiative was designed specifically to bridge these divides. Rather than assume a single trajectory for artificial intelligence or a single theory of risk, the project sought to convene experts from across technical, policy, national security, industry, philanthropic, and civil society communities to examine how AI interacts with global risks under conditions of uncertainty. Rather than consensus, this project focused on developing a more coherent and actionable understanding of how AI may reshape risk through threat, vulnerability, and consequence across domains.

This report emerged from that broader effort. It is, ultimately, a synthesis product: one that reflects recurring themes, tensions, and governance challenges surfaced over the course of this project through cross-domain engagement.

Convening and Engagement Process

Between 2025 and 2026, the AIxGR project convened 6 roundtables, 6 briefings to key offices and policymakers, and a press briefing focused on the relationship between artificial intelligence and global risk. These engagements brought together participants from across the U.S. government, academia, civil society organizations, frontier AI companies, philanthropic organizations, and the broader technical and national security communities. A list of the roundtables and their focus areas is given below.

CONVENING	FOCUS AREA	CONVENING	FOCUS AREA
ROUNDTABLE 1	AI X NUCLEAR COMMAND, CONTROL, AND STRATEGIC STABILITY	ROUNDTABLE 4	AI X MILITARY INTEGRATION AND DECISION-MAKING
ROUNDTABLE 2	AI X BIOSECURITY AND BIOLOGICAL RISK	ROUNDTABLE 5	AGI, FRONTIER SYSTEMS, AND AUTONOMOUS POWER
ROUNDTABLE 3	AI X CYBERSECURITY AND CYBER OPERATIONS	ROUNDTABLE 6	CROSS-DOMAIN AI X GLOBAL RISK INTEGRATION

Across these engagements, participants examined both near- and long-term challenges associated with increasingly capable AI systems. Discussions explored issues including escalation risk, crisis stability, AI-enabled cyber operations, biological uplift, evaluation and testing gaps, human-machine interaction, autonomous systems, information integrity, resilience, and institutional adaptation under competitive pressure.

The initiative also included expert consultation, intergovernmental engagement, and briefings with policymakers and national security stakeholders. We are proud to say we engaged several hundred participants across the AI, global risk, and national security ecosystems during this project.

Importantly, the initiative was intentionally interdisciplinary. Participants often approached AI from fundamentally different assumptions, professional incentives, and theories of risk. Some viewed AI primarily as a powerful but ultimately normal technology requiring careful governance and institutional adaptation. Others emphasized the possibility of more autonomous and transformative systems that may become increasingly difficult to control. Still others remained skeptical of claims surrounding AI capability and argued for greater restraint against hype and speculative forecasting.

Recurring Themes and Related Materials

Several themes consistently emerged across the project's engagements.

Participants emphasized that AI increasingly acts as a cross-domain capability rather than a standalone technology. Its integration into biological research, cyber operations, military systems, intelligence analysis, and critical infrastructure complicates traditional governance models that were built around more discrete technological domains.

Experts also surfaced the growing importance of thoughtfully understanding and building sociotechnical systems. Many of the most significant risks discussed did not stem solely from AI model performance in isolation: rather, those risks emerged due to how AI systems interact with institutions, incentives, workflows, infrastructure, and human operators under conditions of uncertainty and competition.

Testing, evaluation, and institutional adaptation were also front of mind. Across domains, our experts noted that AI capabilities often appear to advance faster than the systems responsible for evaluating, governing, and safely integrating them.

Finally, discussions repeatedly returned to the problem of uncertainty. Participants held diverging perspectives on the pace and trajectory of AI development, but broadly converged on the need for governance approaches that: 1) remain adaptive; 2) are evidence-driven; and 3) are resilient across multiple possible futures.

These themes helped shape both the analytical framework and policy recommendations presented within this report.

Appendix B. Contributors

The Federation of American Scientists and the Future of Life Institute thank the individuals listed below who have provided consent for public acknowledgement of their participation in the roundtables, consultations, briefings, and related engagements associated with the AIxGR project. Affiliations are listed for identification purposes only, and do not imply endorsement of this report or its conclusions.

NAME	TITLE	AFFILIATION
BRIAN ABEYTA	INDEPENDENT CONSULTANT	ABEYTA GPS LLC
JAMES ACTON	CO-DIRECTOR, NUCLEAR POLICY PROGRAM	CARNEGIE ENDOWMENT FOR INTERNATIONAL PEACE
RYAN P. BADMAN	RESEARCH ASSOCIATE	HARVARD MEDICAL SCHOOL
DARIA BAHRAMI	HEAD OF POLICY	DREADNODE
JON BATEMAN	SENIOR FELLOW AND CO-DIRECTOR	TECHNOLOGY AND INTERNATIONAL AFFAIRS PROGRAM CARNEGIE ENDOWMENT FOR INTERNATIONAL PEACE
MARK BEALL	PRESIDENT	AI POLICY NETWORK
DEREK BELLE	ASSOCIATE DIRECTOR, AI & EMERGING TECHNOLOGY INITIATIVE	THE BROOKINGS INSTITUTION
ELIZABETH BODINE-BARON	SENIOR INFORMATION SCIENTIST	RAND
MALO BOURGON	CEO	MACHINE INTELLIGENCE RESEARCH INSTITUTE (MIRI)
DOMINIC BRENNAN	DIRECTOR	INSTITUTE ON GLOBAL CONFLICT AND COOPERATION
CHARLIE BULLOCK	SENIOR RESEARCH FELLOW	INSTITUTE FOR LAW & AI
SARAH R. CARTER	PRINCIPAL	SCIENCE POLICY CONSULTING
BRANDON CORTINO	SENIOR ASSOCIATE FOR NUCLEAR POLICY	THE INSTITUTE FOR SECURITY AND TECHNOLOGY
DONALD L COULTER	MR.	DHS SCIENCE & TECHNOLOGY
ERICH DEVENDORF	AI SECURITY PORTFOLIO LEAD	RAND
TIMOTHY DITTER	SENIOR RESEARCH SCIENTIST	CNA
JEAN DONG	RESEARCH FELLOW	HARVARD KENNEDY SCHOOL
JANET EGAN	SENIOR FELLOW AND DEPUTY DIRECTOR, TECHNOLOGY AND NATIONAL SECURITY	CENTER FOR A NEW AMERICAN SECURITY
ROBERT K. ELDER	PRESIDENT & CEO	OUTRIDER FOUNDATION
LIAM EPSTEIN	RESEARCH ASSISTANT	CENTER FOR A NEW AMERICAN SECURITY
RYAN FEDASIUK	FELLOW, CHINA AND TECHNOLOGY	AMERICAN ENTERPRISE INSTITUTE
MATT FERREN	FELLOW	COUNCIL ON FOREIGN RELATIONS
MATTHEW GENTZEL	NUCLEAR WEAPONS POLICY PROGRAM OFFICER	LONGVIEW PHILANTHROPY
JULIE GEORGE	RESEARCH FELLOW	CENTER FOR SECURITY AND EMERGING TECHNOLOGY (CSET)
ERIC GOMEZ	HORIZON INSTITUTE FELLOW	FEDERATION OF AMERICAN SCIENTISTS
GIGI KWIK GRONVALL	PROFESSOR	JOHNS HOPKINS BLOOMBERG SCHOOL OF PUBLIC HEALTH

NAME	TITLE	AFFILIATION
STEPHANIE GUERRA	SENIOR RESEARCH RESIDENT	RAND
CHAD HEITZENRATER	SR. INFORMATION SCIENTIST	RAND CORPORATION
THOMAS HERNLY	VP, DIGITAL SOLUTIONS ENGINEERING & AI INNOVATION	R&T DEPOSIT SOLUTIONS
REBECCA HERSMAN	SENIOR RESEARCH SCHOLAR	GOVAI
QUENTIN HODGSON	SENIOR RESEARCHER	RAND
MAXIMILIAN HOELL	SENIOR NUCLEAR PROGRAM ASSOCIATE	LONGVIEW PHILANTHROPY
MICHAEL HOROWITZ	DIRECTOR, PERRY WORLD HOUSE	UNIVERSITY OF PENNSYLVANIA
TOM INGLESBY	DIRECTOR	JOHNS HOPKINS CENTER FOR HEALTH SECURITY
KRYSTAL JACKSON	DIRECTOR FOR AI SECURITY	INSTITUTE FOR SECURITY AND TECHNOLOGY
CAROLINE JEANMAIRE	INTERIM DIRECTOR, AI SECURITY POLICY	THE FUTURE SOCIETY
ELIANA JOHNS	SENIOR RESEARCH ASSOCIATE	FEDERATION OF AMERICAN SCIENTISTS
JORDAN KANE	FELLOW	GOVAI
STEVE KELLY	CHIEF TRUST OFFICER	INSTITUTE FOR SECURITY AND TECHNOLOGY
KYLE A KILIAN	SENIOR TECHNICAL ANALYST	RAND
CLARA LANGEVIN	SENIOR MANAGER, AI POLICY	FEDERATION OF AMERICAN SCIENTISTS
SHENG LIN-GIBSON	CHIEF	NIST BIOSYSTEMS AND BIOMATERIALS DIVISION
ANDREW J. LOHN	SENIOR FELLOW	CENTER FOR SECURITY AND EMERGING TECHNOLOGY
AUSTIN LONG	SENIOR FELLOW	MIT CENTER FOR NUCLEAR SECURITY POLICY
CHRIS MESEROLE	EXECUTIVE DIRECTOR	FRONTIER MODEL FORUM
SYLVIA MISHRA	DIRECTOR OF NUCLEAR POLICY	INSTITUTE FOR SECURITY AND TECHNOLOGY
STEVEN MOSS	SENIOR POLICY ADVISOR	NATIONAL SECURITY COMMISSION ON EMERGING BIOTECHNOLOGY
JOSHUA NEW	DIRECTOR OF POLICY	SEEDAI
MICHELLE NIE	VISITING FELLOW, TECHNOLOGY AND NATIONAL SECURITY	CENTER FOR A NEW AMERICAN SECURITY
ABI OLVERA	FOREIGN SERVICE OFFICER (PERSONAL CAPACITY)	DEPT OF STATE (PERSONAL CAPACITY)
CHRISTOPHER J. PARK	PRINCIPAL CONTRARIAN	C.J. PARK LLC
DR. MICHAEL PARKER	ASSISTANT DEAN	GEORGETOWN UNIVERSITY
MATEO PETEL	RESEARCH SCIENTIST	STANFORD UNIVERSITY
DR. JOEL PREDD	DIRECTOR, CENTER FOR THE GEOPOLITICS OF ARTIFICIAL GENERAL INTELLIGENCE	RAND
KRISTIN SCHNEEMAN	SENIOR DIRECTOR, FASTERCURES	MILKEN INSTITUTE
OLIVIA SHOEMAKER	LEAD ADVISOR FOR AI	FRONTIER DESIGN
TRISHA TUCHOLSKI	PROGRAM OFFICER	NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE
VIKRAM VENKATRAM	RESEARCH ANALYST	CENTER FOR SECURITY AND EMERGING TECHNOLOGY (CSET), GEORGETOWN UNIVERSITY
JACQUELINE VITEZNIK	POLICY ANALYST	AMERICANS FOR RESPONSIBLE INNOVATION
DAN WACHTLER	CEO	VIGILIS, INC.
STERLIN WATERS	U.S. PUBLIC POLICY LEAD	CENTER FOR AI RISK MANAGEMENT AND ALIGNMENT

NAME	TITLE	AFFILIATION
BENJAMIN WEINSTEIN-RAUN	SENIOR RESEARCHER	PALISADE RESEARCH
JAIME YASSIF	VICE PRESIDENT	NTI | BIO
TONG ZHAO	SENIOR FELLOW	CARNEGIE ENDOWMENT FOR INTERNATIONAL PEACE

FAS FEDERATION OF AMERICAN SCIENTISTS

About the Federation of American Scientists

The Federation of American Scientists is dedicated to democratizing the policymaking process by working with new and expert voices across the science and technology community, helping to develop actionable policies that can improve the lives of all Americans. For more about the Federation of American Scientists, visit **FAS.org.**