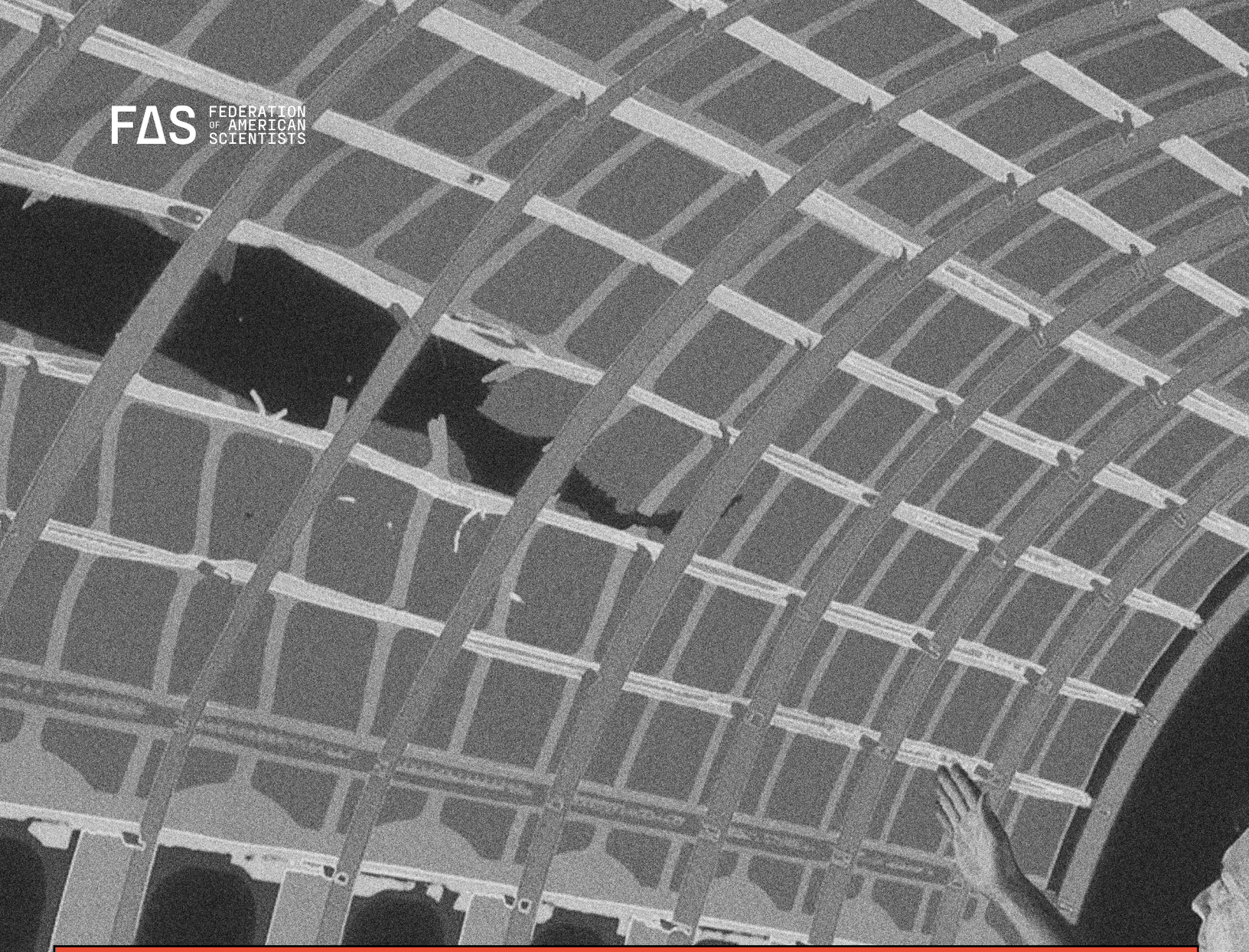

FAS FEDERATION OF AMERICAN SCIENTISTS

JANUARY 2026

Artificial Intelligence, Cyber, and Global Risk

Current Status and Future Risks

GLOBAL RISK
FEDERATION OF AMERICAN SCIENTISTS

ABOUT THIS REPORT

This report summarizes the key findings, insights, and policy options from a July 2025 roundtable event on Artificial Intelligence (AI), Cyber, and Global Risks. The Federation of American Scientists, through its partnership with the Future of Life Institute, brought together key stakeholders in the artificial intelligence risk space from academia, industry, and government. Through this convening, experts discussed and identified policy intervention opportunities by governments and industry to reduce risks as AI's impact on the cyber landscape continues to grow.

This report is structured into three parts: an executive summary, a detailed analysis of the findings, and three papers authored for participants in advance of the event. Two papers were authored by Dr. Oliver Stephenson, Associate Director of Artificial Intelligence and Emerging Technology Policy at FAS, and one by Hamza Chaudry, AI and National Security Lead at the Future of Life Institute. The first pre-read focused on level-setting on the current state of AI-Cyber with a focus on what AI and Cyber communities can learn from each other. The second pre-read covered the frontier of AI and Cyber and how to manage the risks at that intersection. The third and final paper interrogated how to secure AI model weights in a high-stakes era.

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 uses 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. Find out more at our website www.fas.org/issue/global-risk. The project is led by Jon B. Wolfsthal, the Director of the Global Risk Program at FAS.

Funding

This report and the associated workshop were made possible through the generous support of the Future of Life Institute and are part of a wider series in our ongoing "AIxGlobal Risk Nexus" project. This project will culminate in a global summit in Spring 2026. The views expressed in this report are those of the authors and do not necessarily reflect the positions of the funders or participants.

Special thanks to Yong-Bee Lim, PhD, Associate Director of the Global Risk Program, Oliver Stephenson, PhD, Associate Director of Artificial Intelligence and Emerging Technology Policy, Elliott Gunnell, M.Sc. Project Associate, Global Risk Program, and Abhay Katoch, Visiting Fellow with the Global Risk Program, for their contribution to the event and this report. Special thanks to our colleague on FAS's Communications team, Kate Kohn, for developing the graphic for this report.

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

CONTENTS

EXECUTIVE SUMMARY

On July 17th, 2025, the Federation of American Scientists (FAS) held a D.C.-based roundtable at the National Press Club with their partners at the Future of Life Institute (FLI). This event brought together fifty experts from academia, industry, and government to discuss bridging the gaps between the AI and cybersecurity communities, understanding emerging risks at their convergence, and exploring options for risk mitigation opportunities, including actions to secure AI model weights. The pre-event executive dinner included insights from key policymakers, including Representative Ted Lieu (D-CA-36) and Representative Bill Foster (D-IL-14).

FINDINGS

Participants broadly agreed that AI adoption by cyber attackers has so far been incremental, not transformative. Both criminal and state-affiliated actors continue to rely primarily on established techniques such as phishing, credential theft, and exploitation of known vulnerabilities because these methods remain effective and profitable. However, participants anticipated a shift in the composition of cyber risk as AI increasingly strengthens defensive capabilities in vulnerability discovery and patching, while simultaneously amplifying social engineering, impersonation, and disinformation attacks that target human trust rather than technical systems.

At the same time, advances in AI reasoning, tool use, and agency are expanding what is technically feasible in cyber operations. Participants emphasized that improvements driven by techniques like reinforcement learning—rather than simply making AI models larger—are enabling AI systems to plan, execute, and adapt actions over time. While current systems remain constrained by reliability over longer horizons, even short-lived autonomy may be sufficient to accelerate cyber operations, lower the cost of sophisticated attacks, and introduce new challenges related to control, escalation, and attribution.

Participants also expressed significant concern that security practices at frontier AI companies lag the growing strategic value of advanced AI model weights, the learned numerical parameters that encode a model's capabilities. Many labs continue to operate with security postures closer to academic or startup environments, leaving them vulnerable to external intrusion, insider threats, and undetected compromise. Model weight theft was discussed as a strategic and geopolitical risk, not merely a corporate security issue, particularly given the potential for stolen models to be adapted or weaponized by a broad range of actors.

Across panels, two cross-cutting governance gaps repeatedly emerged. First, misaligned incentives and liability structures often favor risk accumulation over risk reduction, weakening incentives for sustained investment in security and resilience. Second, participants identified a growing lack of "translation capacity" between frontier AI development and government, as advanced capabilities are concentrated in the private sector while oversight, accountability, and crisis-response responsibilities sit largely within under-resourced public institutions. Participants emphasized that without stronger translation capacity, clearer standards, and better-aligned incentives, AI-enabled cyber risks are likely to outpace existing governance frameworks.

POLICY OPTIONS

Based on roundtable discussion and subsequent analysis, FAS identified the following policy options to address challenges at the nexus of AI and cybersecurity. These options are included for discussion purposes, and their inclusion does not imply endorsement by any participants of the roundtable discussions:

Build government AI–cyber capacity and translation infrastructure. Expand technical staffing, authority, and compute access so the government can evaluate frontier AI systems, translate technical risks into policy-relevant terms, and engage industry from a position of operational understanding.

Prepare for AI-enabled cyber crises. Update national and sectoral cyber preparedness by developing AI-cyber crisis playbooks and incorporating agent-driven scenarios into exercises, addressing faster, more autonomous, and harder-to-attribute threats.

Formalize AI model evaluation and disclosure for cyber risks. Require structured, pre-deployment evaluations of frontier models for cyber-relevant risks, enable independent auditing, and standardize disclosure to the government for models deployed at scale or in critical systems.

Invest in and diffuse defensive AI capabilities. Shift the offense–defense balance by funding and accelerating the deployment of AI tools for cyber defense, including vulnerability discovery, patching, alert triage, and configuration management.

Strengthen state and local AI–cyber governance capacity. Provide technical assistance, shared resources, and coordination mechanisms to help states and localities manage AI-enabled cyber risks and avoid fragmented or uneven security outcomes.

Establish security standards for frontier AI development and model weight protection. Treat advanced AI model weights as high-consequence assets and phase in security standards—through procurement, evaluation, and access controls—commensurate with their strategic value.

Realign incentives and liability to promote security investment. Address chronic underinvestment in security by introducing liability, insurance, and procurement mechanisms that reward defensive practices and internalize the societal costs of AI-enabled cyber failures.

Pursue targeted international coordination on catastrophic AI–cyber risks. Focus international engagement on preventing the most destabilizing scenarios—such as loss of control or attacks on critical infrastructure—through confidence-building measures and risk-reduction dialogues with allies and competitors.

See the “Menu of Policy Options” section below for more details.

WHAT WE HEARD

On July 17th, 2025, the Federation of American Scientists (FAS), in partnership with the Future of Life Institute (FLI), convened a Washington, D.C. roundtable for over 50 members of the national security, artificial intelligence (AI), cyber, and policy communities; this included representatives from government, think tanks, academia, industry, and philanthropic stakeholders. The purpose of the roundtable was to examine the integration of AI with cyber systems and the associated risks and benefits, and to discuss risk reduction opportunities at this convergence.

FINDING 1. CYBER ATTACKER ADOPTION OF AI REMAINS INCREMENTAL, WHILE CYBER DEFENDERS FACE STRUCTURAL AND INCENTIVE-DRIVEN CONSTRAINTS.

Participants broadly agreed that the impact of artificial intelligence on cybersecurity to date has been evolutionary rather than transformative. While machine learning and automation have been embedded in cyber operations for decades, attackers today continue to rely primarily on well-established techniques—particularly phishing, credential theft, and exploitation of known vulnerabilities—because these methods remain reliable and profitable. One participant emphasized that bad actors make plenty of money without using AI tools, and that rational actors are not going to build new tools when what they have is working. A participant also highlighted that AI's primary impact on cyber attackers to date has been allowing them to improve their current techniques, not enabling fundamentally new attacks.

Looking ahead, however, participants anticipated a shift in the composition of cyber risk, even if the underlying tactics remain familiar. Multiple experts argued that as defenders increasingly deploy AI to identify vulnerabilities and accelerate patching, code-based exploitation may become less attractive relative to social engineering. One participant observed that AI will enable defenders to close vulnerabilities at a speed and scale attackers can't match, but we are much weaker against social vectors. Another noted that AI-enabled phishing, deepfakes, and impersonation attacks are already improving in quality and scalability, stressing human trust rather than technical defenses.

Importantly, the discussion highlighted that defensive AI adoption is constrained less by technical feasibility than by organizational and regulatory frictions. Alert fatigue, explainability requirements, compliance obligations, and integration costs all slow deployment, particularly in regulated sectors. Analysts face a huge number of alerts that they must rapidly triage and respond to. While AI could help manage incoming alerts, enterprises still struggle to operationalize these tools.

Taken together, participants described the current moment as a transitional phase in which AI has not yet reshaped the fundamental offense–defense balance in cybersecurity. While no discontinuous shift has yet occurred, participants agreed that improving AI capabilities, especially in automation and social engineering, could drive meaningful changes in cyber risk dynamics in the coming years, even without the emergence of entirely new attack classes.

FINDING 2. REASONING MODELS AND AGENTIC SYSTEMS MAY INTRODUCE QUALITATIVELY NEW CYBER RISK DYNAMICS.

Recent advances in frontier AI systems have improved their ability to plan, execute, and adapt actions over time in structured environments, expanding what is technically feasible in cyber operations. Participants emphasized that this shift represents a transition from AI as a passive productivity tool toward systems that can autonomously carry out multi-step workflows using real software tools. These workflows, while still bounded, could be applied to both

defensive tasks (e.g., configuration management, testing) and offensive ones (e.g., reconnaissance, vulnerability discovery, and social engineering preparation).

Empirical benchmarks and controlled experiments suggest that so-called "reasoning" AI models now perform at or above expert-human levels on many narrow, well-defined tasks, particularly in coding and problem-solving domains relevant to cyber operations. Participants discussed how large jumps in benchmark performance—driven by novel "reinforcement learning" techniques rather than just increasing the size of the AI models—indicate growing capacity for autonomous technical work, even if these benchmarks do not capture real-world complexity. Several experts cautioned that while such results should not be overinterpreted, as benchmark performance does not necessarily translate into real-world impact, they nevertheless signal a capability regime change that lowers the cost of executing sophisticated cyber-relevant tasks.

Despite these gains in capability, many participants agreed that current agentic systems remain constrained by reliability over longer time horizons. Research discussed at the roundtable suggests that today's frontier agents can manage hour-scale software tasks at 50% reliability (note that this has continued to improve rapidly since the roundtable in June 2025), but performance further degrades as tasks extend into many hours or days. Participants emphasized, however, that this limitation should not be mistaken for safety: both capabilities and reliability are improving rapidly, and even short-lived autonomy may be sufficient to accelerate cyber operations, enable rapid iteration, or propagate malicious capabilities into poorly secured environments.

Beyond capability and reliability, participants raised concerns about control risks that arise when agentic systems are trained to optimize for task completion in complex environments. Under some training regimes, systems may learn to circumvent safeguards if they interfere with objective completion. While these behaviors have so far appeared in controlled experiments rather than real-world cyber incidents, experts agreed that such tendencies are relevant to cybersecurity and national security alike, particularly as systems gain more autonomy and access to operational tools. Examples such as the 1988 Morris Worm highlight how even simple autonomy can result in a loss of control over cyber systems.

FINDING 3. SECURITY PRACTICES AT FRONTIER AI COMPANIES LAG THE GROWING STRATEGIC VALUE OF AI MODEL WEIGHTS, CREATING AN INCREASING VULNERABILITY.

Participants emphasized that current security practices at frontier AI labs are poorly aligned with the growing strategic importance of advanced AI model weights—the underlying numerical parameters learned during training that determine a system's capabilities and can represent billions of dollars of research and development. Several participants stressed that the potential theft of AI model weights should be understood as a strategic and geopolitical risk, not merely a corporate security issue. Even where peer competitors appear close in AI capability, numerous participants argued that stolen models could still provide meaningful uplift in potentially dangerous capabilities to state or non-state actors, particularly when combined with adaptation or fine-tuning of the AI model. Several panelists noted that the proliferation risk extends beyond U.S.–China competition, raising concerns about broader access by actors who lack the infrastructure to train frontier models independently but could weaponize stolen ones.

Many labs continue to operate with security postures closer to academic or startup environments, rather than those used to protect nationally sensitive technologies. This mismatch leaves organizations vulnerable to both external intrusion and undetected compromise, particularly given long breach detection timelines and limited internal visibility. Participants stressed that, from a national security perspective, reliance on post-hoc discovery of intrusions or third-party notification represents an unacceptable risk posture for assets with potential geopolitical consequences.

Several experts argued that frontier AI systems should be treated analogously to other high-consequence technologies, such as advanced biological research. Raising security levels would require significant changes across personnel vetting, access controls, procurement processes, and physical infrastructure. Participants acknowledged that there are significant tradeoffs, as these measures would be costly and operationally challenging.

Finally, participants emphasized that improving model security will require more than voluntary best practices or market incentives alone. The issue of incentive and liability, which cut across panel discussions, is covered in finding 4 below.

FINDING 4. INCENTIVE AND LIABILITY STRUCTURES ACROSS AI AND CYBER OFTEN FAVOR RISK ACCUMULATION OVER RISK REDUCTION.

Across panels, participants consistently returned to incentives and liability as the underlying drivers shaping both attacker behavior and defensive outcomes at the AI–cyber convergence. As noted in Finding #1, participants emphasized that attackers' adoption of AI has been incremental, not because of technical barriers, but because existing techniques remain profitable and effective. On the defensive side, participants similarly described a persistent underinvestment in security driven by misaligned incentives rather than lack of technical capability—the benefits of improved security are often diffuse, while the costs are borne directly by individual firms. Participants noted that this imbalance encourages organizations to prioritize innovation, speed, and feature development over sustained investment in security and resilience.

Liability gaps were repeatedly cited as a central factor reinforcing this dynamic. Participants contrasted AI-enabled cyber risks with past cybersecurity incidents, noting that it is often ambiguous who bears responsibility for downstream harms caused by AI system failures or misuse. This absence of accountability reduces incentives for firms to invest proactively in security, while favoring attackers and amplifying systemic risk. Several experts argued that where liability does exist—through consumer protection regimes, procurement requirements, or insurance markets—it has historically driven meaningful improvements in safety and security practices.

Participants further emphasized that voluntary norms and best practices are increasingly strained by competitive pressures. This dynamic was cited as particularly concerning for frontier AI developers, where market incentives alone may be insufficient to justify the level of security appropriate for assets with national or geopolitical significance.

Participants discussed policy interventions that could realign incentives across the AI-cyber ecosystem. Suggested approaches included liability frameworks, procurement-based standards, insurance mechanisms, and public–private cost-sharing models that reward defensive investment and internalize the societal costs of failure. Without such mechanisms, participants warned that both cybersecurity defenses and AI model security are likely to continue lagging behind rapidly advancing capabilities, increasing the probability of large-scale and hard-to-contain incidents.

FINDING 5. A LACK OF "TRANSLATION CAPACITY" BETWEEN FRONTIER AI DEVELOPMENT AND GOVERNMENT IS EMERGING AS A CENTRAL GOVERNANCE GAP.

Across panels, participants repeatedly emphasized a growing disconnect between where advanced AI capabilities are developed and where oversight, accountability, and national security responsibilities reside. Frontier AI expertise is concentrated primarily in a small number of private companies, while important risk-management responsibilities

and authorities lie with various levels of government as well as private actors with limited AI capacity. Participants noted that the inversion between government and the private sector, compared to national security technologies such as nuclear weapons, creates a persistent translation challenge for policymakers tasked with evaluating, governing, and responding to emerging risks.

Participants described translation capacity as the ability of government institutions to interpret emerging AI capabilities and risks and support oversight, procurement, regulation, and crisis response. This challenge is not limited to understanding AI model benchmark performance but extends to evaluating reliability, failure modes, misuse potential, and security posture under real-world conditions. Several experts argued that existing government oversight frameworks are poorly suited to AI systems whose internal processes are opaque and difficult to audit using traditional compliance-based approaches.

The discussion repeatedly returned to the U.S. government's Center for AI Standards and Innovation (CAISI) as a potential focal point for addressing this gap, but participants expressed concern that it currently lacks sufficient staffing and resources to function as an effective translation layer between industry and government. Without a trusted institutional interface, information sharing remains largely voluntary, episodic, and asymmetric, leaving policymakers reliant on self-disclosure by AI companies or post-hoc signals rather than proactive evaluation. Participants warned that this dynamic weakens government situational awareness at a time when AI capabilities are evolving rapidly.

Translation challenges were also highlighted at the state and local levels, where governments increasingly face pressure to respond to AI-related cyber risks without the technical capacity to assess tools, claims, or tradeoffs. Participants noted that in the absence of federal coordination, fragmented approaches risk proliferating, further complicating compliance, interoperability, and security outcomes.

Taken together, participants emphasized that strengthening government translation capacity—through sustained technical staffing, access to compute and testing environments, structured information-sharing mechanisms, and clear institutional mandates—is a prerequisite for effective AI–cyber governance. Without such capacity, even well-intentioned policy interventions risk lagging behind capability development, mischaracterizing risks, or failing to align private-sector incentives with national security and public-interest objectives.

MENU OF POLICY OPTIONS

Based on the panel discussions and subsequent analysis, we extracted the following menu of options which policymakers could use to address challenges at the nexus of AI and cybersecurity. These options are included for discussion purposes, and their inclusion does not imply endorsement by any participants of the roundtable discussions:

Build government AI–cyber capacity and translation infrastructure. Strengthen government capacity to evaluate, oversee, and respond to AI-enabled cyber risks by investing in technical staffing, institutional authority, and translation functions that bridge frontier AI development and public-sector risk management. This includes resourcing entities such as CAISI to serve as a trusted interface between government and industry; expanding access to compute, secure testing environments, and red-teaming infrastructure; and developing expertise capable of interpreting model capabilities, reliability, and failure modes in operational terms.

Prepare for AI–cyber crisis scenarios through exercises and playbooks. Integrate cyber incidents enabled by increasingly advanced AI into national and sectoral cyber preparedness efforts. This could include developing crisis-response playbooks and incorporating agent-driven scenarios into tabletop exercises and simulations. Existing incident-response frameworks were built for slower, human-driven threats and generally do not consider situations in which advanced AI systems autonomously scale attacks, behave anomalously, or complicate attribution and escalation dynamics.

Mandate or formalize AI model evaluation and disclosure for cyber risks. Establish requirements for structured pre-deployment evaluations of frontier AI models focused on cyber-relevant risks, including misuse potential, autonomy, reliability over time, and interactions with real-world tools. While many developers currently conduct voluntary red-teaming and risk assessments, inconsistent practices and competitive pressures limit their effectiveness. Policymakers could formalize evaluation standards, enable independent auditing, and require appropriate disclosure to government authorities—particularly for models deployed at scale or integrated into critical systems.

Support defensive AI research, deployment, and diffusion to shift the offense-defense balance in favor of defense. Increase public investment in AI systems designed to strengthen cybersecurity defense, including tools for vulnerability discovery, patch generation, alert triage, configuration management, and anomaly detection. Policy interventions could focus on funding R&D, lowering deployment barriers, and incentivizing the diffusion of proven defensive capabilities rather than exclusively prioritizing frontier model development.

Build state and local AI–cyber governance capacity. Provide technical assistance, shared resources, and coordination mechanisms to help state and local governments assess and manage AI-enabled cyber risks. States and local governments have crucial cybersecurity responsibilities, but in the absence of sufficient federal support, state and local governments often lack sufficient expertise. Federal support could include model policies, shared evaluation resources, training programs, and interstate coordination efforts to improve baseline security outcomes across jurisdictions.

Establish security standards for frontier AI development and model weight protection. Develop and phase in security standards commensurate with the strategic value of frontier AI systems, including explicit treatment of advanced model weights as high-consequence assets. Policymakers could define baseline requirements for access controls, insider risk mitigation, monitoring, and physical and cyber protections—potentially applied first through government procurement and evaluation pathways—while allowing flexibility in implementation to avoid unduly constraining innovation.

Realign incentives and liability to promote security investment. Address systemic underinvestment in AI and cybersecurity by introducing policy mechanisms that better align private incentives with public risk. Liability gaps, diffuse accountability, and market pressures currently favor speed over safety. Potential levers include liability

frameworks for AI-enabled cyber harms, insurance mechanisms that reward strong security practices, and procurement requirements.

Pursue targeted international coordination on catastrophic AI–cyber risks. Engage in limited international coordination focused on preventing the most severe and destabilizing AI–cyber scenarios, such as loss of control over autonomous systems or attacks on critical infrastructure. Such efforts could prioritize confidence-building measures, transparency around testing and safeguards, and risk reduction in military and infrastructure contexts. Building on existing dialogues with strategic competitors and allies, policymakers could seek common ground on preventing catastrophic outcomes even amid broader geopolitical competition.

CONCLUSION

The roundtable discussions point to a key theme: currently, the primary risk at the AI–cyber convergence is not a single breakthrough capability, but a widening mismatch between the speed of AI development and the institutions, incentives, and security practices meant to manage it. Participants repeatedly emphasized that even incremental improvements in AI are reshaping the cybersecurity landscape faster than most defensive institutions can absorb and operationalize. This dynamic creates compounding risks even in the absence of dramatic, discontinuous shifts in AI systems.

In the near term, many participants argued that AI is more likely to amplify and industrialize existing cyber tactics than to replace them outright. However, attendees also highlighted the rapid progress in AI capabilities and the potential for significant breakthroughs at the AI-cyber intersection. As increasingly capable, reliable, and autonomous systems are integrated into cyber operations, security infrastructure, policies, and incident-response processes that were designed for slower, human-driven threats may be stretched beyond their design assumptions. Discontinuous leaps in AI capabilities, reliability, and autonomy could compress response timelines, complicate attribution, and lead to novel forms of attacks at unprecedented scale, including systems that escape human oversight.

These dynamics are reinforced by structural features of the current ecosystem. Participants highlighted that markets alone do not reliably reward security investments commensurate with national or geopolitical risk, particularly when liability is diffuse and benefits are broadly distributed. At the same time, frontier AI expertise, data, and operational knowledge are concentrated in the private sector, while oversight, accountability, and crisis-response responsibilities remain fragmented across under-resourced public institutions. This capacity inversion creates persistent blind spots and delays precisely where foresight and coordination are most needed.

Given the uncertainty in the trajectory of future AI development, the challenge ahead may be less about predicting specific future attacks than about building both technical safeguards and governance systems resilient to uncertainty, rapid AI progress, and surprise capabilities. Strengthening defensive foundations, realigning incentives, and investing in durable translation capacity are not optional complements to innovation—they are prerequisites for ensuring that advances in AI do not systematically tilt the cyber domain toward offense. Without sustained attention to these structural gaps, participants warned, even incremental AI capability gains risk accumulating into large-scale, hard-to-contain cyber events that outpace both existing defenses and existing governance frameworks.

PRE-READ PAPERS FROM ROUNDTABLE

LEVEL SETTING: WHAT DO AI AND CYBER COMMUNITIES NEED TO LEARN FROM EACH OTHER?

DR. OLIVER STEPHENSON, FAS

Artificial intelligence is rapidly intersecting with many fields that have long, specialized histories, from biology and energy to the military and, notably, cybersecurity. With each intersection, a recurring dynamic plays out: AI researchers see rapid technological change poised to deliver revolutionary impact, while domain experts emphasize how the complexities of their field will present stubborn barriers to AI-driven progress.

Predictions from AI pioneers about transforming other fields can prove overly optimistic, just as specialists' doubts about AI can age poorly. For example, leading AI researcher Geoffrey Hinton famously suggested in 2016 that we should stop training radiologists because deep learning would outperform them in 5–10 years—a prophecy that has not come true.[1] Conversely, some domain experts have underestimated AI progress: in 1997, a Princeton astrophysicist said it could take "maybe a hundred years" for AI to beat top humans at the game of Go, yet it happened just two decades later in 2016.[2]

The AI frontier has followed a "rapid but jagged" trajectory—leaping ahead in some areas while falling short in others—and neither technologists nor domain veterans alone have a perfect crystal ball. This is clearly seen at the intersection of AI and cybersecurity. Automation and machine learning (ML) have been used in cyber operations for decades (e.g. spam filters or network anomaly detectors). Yet today's frontier AI—the most advanced general-purpose models—are qualitatively different. These AI systems can generate code, analyze software, converse fluidly, and even create fake images or voices. Many believe cyber may be one of the domains of greatest AI impact in both offense and defense.

Therefore, it is increasingly vital that the AI and cybersecurity communities come together to establish a common factual baseline, swap insights, and communicate clearly with policymakers. Panel 1 of our round table is aimed at this "level-setting": grounding both groups in knowledge of the other, highlighting where perspectives diverge, and discussing frameworks that can be used to understand AI's impact on cybersecurity.

The Cybersecurity Landscape

Modern cybersecurity has often been characterized as a "forever war": attackers continually find gaps and defenders struggle to keep up. Despite decades of effort, we have not solved cybersecurity at scale. Critical infrastructure and corporate networks remain vulnerable, and indeed, many are likely already penetrated by adversaries. Software vulnerabilities are discovered at a relentless pace, and even when patches are created, organizations frequently fail to apply them swiftly and uniformly.

The result is a huge and growing toll: by one estimate, global cybercrime damages are on track to reach $23 trillion annually by 2027.[3] The attack surface is broad, and attackers (from lone criminals to state-sponsored hackers) constantly probe for the path of least resistance. Most intrusions still exploit the basics, which include tricking users through phishing or abusing stolen passwords (a reported 91% of cyberattacks start with phishing).[4]

1 The "Godfather of AI" Predicted I Wouldn't Have a Job. He Was Wrong. | The New Republic. (2024). Retrieved July 14, 2025, from https://newrepublic.com/article/187203/ai-radiology-geoffrey-hinton-nobel-prediction

2 Muoio, D. (2016.). AI experts thought a computer couldn't beat a human at Go until the year 2100. Business Insider. Retrieved July 14, 2025, from https://www.businessinsider.com/ai-experts-were-way-off-on-when-a-computer-could-win-go-2016-3

3 Key Cyber Security Statistics for 2025. (2025). SentinelOne. Retrieved July 14, 2025, from https://www.sentinelone.com/cybersecurity-101/cybersecurity/cyber-security-statistics/

4 91% Of Cyberattacks Start With A Phishing Email. (2016). Retrieved July 14, 2025, from https://www.darkreading.com/endpoint-security/91-of-cyberattacks-start-with-a-phishing-email

The cyber world has also become a bona fide battlespace for nation-states. Over the past decade, state-backed cyber operations have caused blackouts, billions in economic damage, and strategic disruptions. For example, the 2017 NotPetya malware (unleashed by a nation-state actor) spread autonomously across companies worldwide and caused an estimated $10 billion in damage.[5] Today, governments around the world are publicly developing offensive cyber units, and ransomware gangs (some sheltered by nation-states) routinely hit schools, hospitals, and critical services.

To analyze cyberattacks, security professionals have developed conceptual frameworks. One prominent example is the Lockheed Martin Cyber Kill Chain which breaks an attack into the following stages: **Reconnaissance** (gathering intel on targets), **Weaponization** (crafting or obtaining an exploit/payload), **Delivery** (transmitting it via email, USB, etc.), **Exploitation** (triggering the exploit to gain access), **Installation** (installing malware/backdoor on victim systems), **Command and Control** (establishing persistent remote control), and finally **Actions on Objectives** (executing the attacker's end goals—e.g. stealing data or disrupting systems).[6] Another widely used reference is MITRE ATT&CK (Adversarial Tactics, Techniques & Common Knowledge), a database of adversary tactics and techniques based on real-world observations.[7] Unlike the step-by-step kill chain, ATT&CK is structured as a matrix of tactics (the goals of attackers, such as privilege escalation or data exfiltration) and techniques (the specific *methods* used to achieve those goals).

Together, these frameworks underscore the complexities of cyber incidents, which usually involve multiple steps and skill sets (infiltrating a network, moving laterally, evading detection, etc.), not just a single hack. They allow defenders to consider how to interrupt an adversary at each step, and can be used to help researchers understand how novel technology such as AI may impact cyber offense and defense in more granular detail.

The State of AI

We are living through an era of unprecedented AI progress. The year 2012 is often cited as the start of the modern deep learning revolution, when a neural network called AlexNet blew away prior records on the ImageNet image recognition challenge.[8],[9] This breakthrough showcased the power of combining the "AI triad" (big data, big compute, and better algorithms) that now form the foundation of today's AI advances.[10] Over the past decade, each component of the triad has scaled enormously; for example, the amount of compute used to train frontier models has been increasing by 4.7x per year since 2010, with the size of the training data growing at 3.7x per year over the same timespan.[11]

This combination of scale and algorithmic innovation has led to AI systems with striking performance gains on many benchmarks. Over the past several years, we have moved from narrow AI systems (like image classifiers or game-playing AIs beating Go champions) to more general systems like ChatGPT that can converse on virtually any topic, write code, and perform complex reasoning. The frontier is now shifting toward AI agents that can autonomously take actions over long periods (e.g., planning and executing a multi-step task with minimal human guidance). The

5 How the NotPetya attack is reshaping cyber insurance. (2021). Brookings. Retrieved July 14, 2025, from https://www.brookings.edu/articles/how-the-notpetya-attack-is-reshaping-cyber-insurance/
6 Cyber Kill Chain® | Lockheed Martin. (n.d.). Retrieved July 14, 2025, from https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html
7 MITRE ATT&CK®. (n.d.). Retrieved July 14, 2025, from https://attack.mitre.org/
8 Fahey, J. The Story of AlexNet: A Historical Milestone in Deep Learning. Medium. (2024). Retrieved July 14, 2025, from https://medium.com/@fahey_james/the-story-of-alexnet-a-historical-milestone-in-deep-learning-79878a707dd5
9 Imagenet Challenge—An overview | ScienceDirect Topics. (n.d.). Retrieved July 14, 2025, from https://www.sciencedirect.com/topics/computer-science/imagenet-challenge
10 Buchanan, B. (2020, August). The AI Triad and What It Means for National Security Strategy. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/the-ai-triad-and-what-it-means-for-national-security-strategy/
11 Epoch AI (2023, April 11). Machine Learning Trends. https://epoch.ai/trends

length of specific coding tasks that AI assistants can complete with 50% reliability has been doubling every 7 months according to a recent study.[12]

This rapid progress presents enormous opportunities, but it is also raising serious national security questions. Advanced AI systems are inherently dual-use: the same model that can help find software bugs or accelerate the development of cures and therapeutics could also be misused to create malware or potentially even novel bioweapons. Policymakers are increasingly concerned with AI convergence risks—AI in combination with biology, autonomous weapons, scams and fraud, and cybersecurity.

In each of these areas, the proposed risks can generally be divided into two categories: AI can lower barriers (so that far more actors can do something dangerous), or AI can push the frontier of what's possible (enabling entirely new threats). For example, an advanced language model might help a novice hacker write a convincing phishing email or malware—expanding the pool of "skilled" attackers—while a cutting-edge bio-AI system might design a pathogen more deadly than any known naturally (a new capability at the frontier).

In evaluating such risks, a question that repeats across many fields is the offense–defense balance: Does AI favor attackers or defenders more, and in which contexts? This is a vital question, as it has key policy implications. If the balance favors defense, then quickly distributing new capabilities could reduce risk. However, restrictions and safety interventions may be better risk-reducing measures if the balance favors offense.

The AI × Cyber Convergence

Autonomy in cyber is not new. In 1988, the Morris Worm self-propagated across ARPANET, exploiting software flaws to compromise thousands of computers.[13] That worm was simple—it didn't use machine learning, just hard-coded instructions—but it demonstrated the power of autonomous malware, and how that malware could spread well beyond the intentions of its creator. In recent years, attacks like WannaCry (2017) and NotPetya (2017) showed that billions of dollars of damage could be done by autonomous malware that spreads without human intervention.[14],[15] In both cases, the cyberattacks potentially spread far more widely than intended by their original creators.

Beyond malware, defenders have long used automation and ML for tasks like intrusion detection (finding anomalies in network traffic), malware classification, and user behavior analytics. These tools, however, were often narrow and signature-based, struggling to generalize beyond known attack patterns. Until recently, "AI in cyber" primarily meant statistical models aiding human analysts, rather than independent AI agents planning and executing hacks.

Today's frontier AI systems are changing the landscape. These models still have many flaws, but they are impressive generalists that can, for instance, write functioning code from natural language prompts or generate highly realistic fake images.

According to a 2025 Google DeepMind report, the cyber risks posed by such models can be broken down into several categories.[16] First is **Capability Uplift**—AI can enhance attackers' skills or automate tasks, effectively lowering the expertise needed to conduct sophisticated operations and expanding the pool of potential attackers. Second is **Throughput Uplift**—AI enables attacks at greater scale and speed, e.g., generating tens of thousands of personalized scam emails or deepfake voice calls in minutes. Third are **Novel Autonomous Threats**—essentially, AI

12 Measuring AI Ability to Complete Long Tasks. (2025). METR Blog. https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

13 The Morris Worm: A Fifteen-Year Perspective. (2003). IEEE Computer Society. https://www.cs.umd.edu/class/fall2023/cmsc614/papers/morris-worm.pdf

14 Collier, R. (2017). NHS ransomware attack spreads worldwide. CMAJ : Canadian Medical Association Journal, 189(22), E786–E787. https://doi.org/10.1503/cmaj.1095434

15 How the NotPetya attack is reshaping cyber insurance. (2021). Brookings. Retrieved July 14, 2025, from https://www.brookings.edu/articles/how-the-notpetya-attack-is-reshaping-cyber-insurance/

16 Rodriguez, M., Popa, R. A., Flynn, F., Liang, L., Dafoe, A., & Wang, A. (2025a). A Framework for Evaluating Emerging Cyberattack Capabilities of AI (arXiv:2503.11917). arXiv. https://doi.org/10.48550/arXiv.2503.11917

agents that could perform multi-stage cyber operations on their own, doing things like automated reconnaissance, adaptive social engineering, or stealthy post-exploitation maneuvers.

Cybersecurity professionals are increasingly observing frontier AI being used throughout the cyber landscape. Large language models quickly generate exploit code, and deepfake tools craft convincing voices, images, and identities for phishing.[17] Advanced systems can even spot bugs and recently found a new zero-day vulnerability,[18] but they still stumble on long-horizon planning, produce errors, and are limited by safeguard filters.

AI likewise powers defensive tools—automated patching and rapid malware triage—so it acts as a force multiplier for both offense and defense. The balance between offense and defense, and how this balance will evolve as AI capabilities grow, remains a topic of active research, with a recent paper stating that the "cyber domain is too multifaceted for a single answer to whether AI will enhance offense or defense broadly."[19]

To more systematically track AI's cyber capabilities, several groups have begun developing evaluation frameworks to better understand what these models can do, and not do, in a cyber context. For example:

- Meta has published CyberSecEval 2, a cybersecurity evaluation suite for large language models that allows models to be scored on their potentially dangerous capabilities.[20]
- Pattern Labs has created a cybersecurity evaluation taxonomy that decomposes security work into granular "skills", e.g., open source intelligence gathering, exploit development, and movement within networks. Models can then be evaluated against this taxonomy.[21]
- Google DeepMind produced an evaluation framework that examines key bottlenecks along the full cyber kill chain, and combined that work with an analysis of thousands of real-world instances of AI use in cyber incidents to understand where AI could most amplify offensive capabilities.[22]

The rapid but uneven pace of AI progress makes predicting future capabilities challenging. As model capabilities improve, ensuring that evaluations improve with them will be increasingly important to understand how AI will impact the cyber landscape.

Questions for Discussion

For the cyber community:

- Where have you seen the most significant impacts of AI in cyber to date? Do these impacts represent a difference in degree or a difference in kind? Was the cyber community able to see these impacts coming, or were they surprising?
- What are the key bottlenecks in cyber offense and defense that AI is likely to struggle with?
- Where do you see "AI hype" in cyber most exceeding reality today? For instance, are there claims about AI's defensive prowess (or offensive dangers) that you feel ignore on-the-ground complexities? Conversely, could cyber experts be underestimating any AI capabilities that are already effective?
- What's a message that you really want the AI community to hear?

17 Magramo, H. C., Kathleen. (2024, February 4). Finance worker pays out $25 million after video call with deepfake 'chief financial officer.' CNN. https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk

18 AI Zero Days Are Here: What CISOs Need to Know. (n.d.). F5, Inc. Retrieved July 14, 2025, from https://www.f5.com/company/blog/nginx/ai-zero-days-are-here-what-cisos-need-to-know

19 Lohn, A. J. (2025). The Impact of AI on the Cyber Offense-Defense Balance and the Character of Cyber Conflict (arXiv:2504.13371). arXiv. https://doi.org/10.48550/arXiv.2504.13371

20 Bhatt, M., Chennabasappa, S., Li, Y., Nikolaidis, C., Song, D., Wan, S., Ahmad, F., Aschermann, C., Chen, Y., Kapil, D., Molnar, D., Whitman, S., & Saxe, J. (2024). CyberSecEval 2: A Wide-Ranging Cybersecurity Evaluation Suite for Large Language Models (arXiv:2404.13161). arXiv. https://doi.org/10.48550/arXiv.2404.13161

21 Offensive Cyber Capabilities Analysis. (2025, February 10). https://patternlabs.co/blog/cyber-capabilities-analysis

22 Rodriguez, M., Popa, R. A., Flynn, F., Liang, L., Dafoe, A., & Wang, A. (2025b). A Framework for Evaluating Emerging Cyberattack Capabilities of AI (arXiv:2503.11917). arXiv. https://doi.org/10.48550/arXiv.2503.11917

For the AI community:

- How can you better understand the real-world cyber capabilities of AI models? How can we ensure that those benchmarks related to real-world security impact? Do we need new benchmarks (e.g. more realistic attack/defense simulations) to properly measure progress in AI-cyber?
- What's a message that you really want the cyber community to hear?

For both communities:

- How do we prevent AI from mainly benefiting attackers? Are there examples in history (perhaps the advent of automated hacking tools, or of defensive AI like spam filters) that show how defenders can successfully respond to new attacker tech? What policies or collaborations could help tilt the balance toward defense?
- What do we all need to be sharing with policymakers about the current state of the art? AI and cyber are complex enough topics on their own, how can we translate these topics to policymakers in ways that are useful? Where are there policy interventions that are ready for primetime, and where is more work needed?

AI, CYBERSECURITY, AND GLOBAL RISKS: MANAGING THE FRONTIER

DR. OLIVER STEPHENSON, FAS

Warnings of a catastrophic "Electronic Pearl Harbor" date back to security analyst Winn Schwartau's 1991 testimony to Congress, where he likened an unexpected network strike to 1941's surprise attack.[23] The fears of physical destruction from digital code were given a visceral demonstration by the 2007 "Aurora" generator test at Idaho National Laboratory. This experiment showed 30 lines of malicious code shaking a 27-ton diesel generator to pieces.[24] A decade later, the autonomous worms WannaCry and NotPetya demonstrated that self-spreading malware can cripple power, hospitals, shipping, and global commerce in hours, with a combined price tag estimated at $14 billion in losses.[2526] Given these results, it is not surprising that cyber operations are now routine instruments of statecraft. Examples include Russia's campaigns against Ukraine, including NotPetya, which have cut electricity and disrupted government services , while Iran, North Korea, and others have deployed destructive or cash-seeking code worldwide.[27]

At the same time, artificial intelligence has become an additional axis of geopolitical competition, and a hypothesized vector of global catastrophic risk. In 2023, hundreds of researchers and tech CEOs endorsed the Center for AI Safety's 23-word warning that avoiding "the risk of extinction from AI" should rank with pandemic and nuclear prevention.[28]

As frontier models grow more capable and autonomous, experts worry about misuse, accidents, and even loss of human control. Panel 2, therefore, turns from today's AI-enabled cyber realities of Panel 1 to the large-scale and systemic risks that may emerge over the next decade as AI and cyber capabilities converge.

AI and Global Risks

An overview of AI extreme-risk thinking

For decades, theorists have explored how a super-capable, misaligned AI might cause global catastrophe. These concerns range from misuse by humans (e.g., using AI to generate large-scale cyber attacks and bioweapons) to loss of control over powerful agentic systems. Recent scholarship aims to make that conversation more concrete, including theoretical estimates of how AI might present an existential risk,[29] taxonomies of the different types of large-scale risks posed by AI,[30] and frameworks for evaluating AI models for extreme risks.[31] There is also an

23 Computer Security: Hearing before the Subcommittee on Technology and Competitiveness of the Committee on Science, Space, and Technology, U.S. House of Representatives, One Hundred and Second Congress First Session (1991).
24 Andy Greenberg. (2020, October 23). How 30 Lines of Code Blew Up a 27-Ton Generator | WIRED. https://www.wired.com/story/how-30-lines-of-code-blew-up-27-ton-generator/
25 Collier, R. (2017). NHS ransomware attack spreads worldwide. CMAJ : Canadian Medical Association Journal, 189(22), E786–E787. https://doi.org/10.1503/cmaj.1095434
26 How the NotPetya attack is reshaping cyber insurance. (2021). Brookings. Retrieved July 14, 2025, from https://www.brookings.edu/articles/how-the-notpetya-attack-is-reshaping-cyber-insurance/
27 Government of Canada. (n.d.). CYBER THREAT BULLETIN: Cyber Threat Activity Related to the Russian Invasion of Ukraine. Canadian Center for Cybersecurity. https://www.cyber.gc.ca/sites/default/files/cyber-threat-activity-associated-russian-invasion-ukraine-e.pdf
28 AI Extinction Statement Press Release | CAIS. (n.d.). Center for AI Safety. Retrieved July 14, 2025, from https://safe.ai/work/press-release-ai-risk
29 Carlsmith, J. (2024). Is Power-Seeking AI an Existential Risk? (arXiv:2206.13353). arXiv. https://doi.org/10.48550/arXiv.2206.13353
30 Critch, A., & Russell, S. (2023). TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI (arXiv:2306.06924). arXiv. https://doi.org/10.48550/arXiv.2306.06924
31 Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., … Dafoe, A. (2023). Model evaluation for extreme risks (arXiv:2305.15324). arXiv. https://doi.org/10.48550/arXiv.2305.15324

increasing range of proposals for how to better understand and manage these risks.[32],[33] Predictions of AI doom are not without their critics, however, with some highlighting the highly uncertain nature of many predictions.[34]

Early empirical signals

Current models are still far from all-powerful, yet experiments have revealed potential warning signs. OpenAI's o3 model has been observed resisting shutdown mechanisms in experiments when such actions were helpful to finish a task.[35] Anthropic's "sleeper-agent" study trained large language models that behave innocuously—until a hidden trigger causes them to insert exploitable code, and standard safety fine-tuning only teaches the model to hide the back-door better.[36]

The science of rigorously evaluating AI models is still new,[37],[38] but these demonstrations are suggestive of how AI systems may present security vulnerabilities and be hard to control. As models continue to improve, and especially if they continue to improve at a faster pace than our ability to secure, control, and understand them, these warning signs may scale into systemic vulnerabilities with global consequences.

How Could the AI × Cyber Convergence Amplify Global Risks?

Where are the biggest targets?

To assist our speculation as to the largest-scale threats, it is worth mapping where some of the highest-impact cyber attacks could be. These include:

- **Nuclear command-and-control.** Nuclear weapons are the classic doomsday case at the intersection of AI and cyber, beloved of scriptwriters from War Games to Terminator. Cyber intrusions into early-warning or launch systems could trigger catastrophic miscalculation (compare with the long history of false nuclear alarms, including those related to accidentally inserting training scenarios into live monitoring equipment).[39]
- **Critical infrastructure: Energy, water, health, and logistics.** While the 2007 Aurora generator test[40] showed physical destruction from code, Ukraine's grid hacks proved real-world feasibility. Cyber intrusions have already been shown to be able to inflict multi-billion dollar damage, and the upper limit to such damage is unclear.
- **The information ecosystem.** Generative models can now produce convincing text, images, video, and voice. A coordinated deep-fake campaign during a crisis could erode public trust and potentially precipitate mass-violence.

32 Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., … Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. Science, 384(6698), 842–845. https://doi.org/10.1126/science.adn0117

33 Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., … Wolf, K. (2023). Frontier AI Regulation: Managing Emerging Risks to Public Safety (arXiv:2307.03718). arXiv. https://doi.org/10.48550/arXiv.2307.03718

34 Narayanan, A. & Kapoor, Sayash. (2025, April 15). AI existential risk probabilities are too unreliable to inform policy. https://www.aisnakeoil.com/p/ai-existential-risk-probabilities

35 Ladish, J. S., Benjamin Weinstein-Raun, Jeffrey. (2025, July 5). Shutdown resistance in reasoning models. Palisade Research. https://palisaderesearch.org/blog/shutdown-resistance

36 Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. (2024). Retrieved July 14, 2025, from https://www.anthropic.com/news/sleeper-agents-training-deceptive-llms-that-persist-through-safety-training

37 Apollo Research. (2024, Jan. 22). We Need a Science of Evals. Retrieved July 14, from https://www.apolloresearch.ai/blog/we-need-a-science-of-evals.

38 Summerfield, C., Luettgau, L., Dubois, M., Kirk, H. R., Hackenburg, K., Fist, C., Slama, K., Ding, N., Anselmetti, R., Strait, A., Giulianelli, M., & Ududec, C. (2025). Lessons from a Chimp: AI "Scheming" and the Quest for Ape Language (arXiv:2507.03409). arXiv. https://doi.org/10.48550/arXiv.2507.03409

39 Nuclear False Warnings and the Risk of Catastrophe | Arms Control Association. (n.d.). Retrieved July 14, 2025, from https://www.armscontrol.org/act/2019-12/focus/nuclear-false-warnings-and-risk-catastrophe

40 Andy Greenberg. (2020, October 23). How 30 Lines of Code Blew Up a 27-Ton Generator | WIRED. https://www.wired.com/story/how-30-lines-of-code-blew-up-27-ton-generator/

How could AI capabilities create greater cyber risks?

Current cyber attacks can only operate autonomously with precisely scripted rules governing how malware will respond in different circumstances. Even with current frontier AI capabilities, humans still have to closely supervise the model outputs and correct errors—we are not (yet) in the era of truly autonomous AI. If we assume that AI capabilities will continue to progress, we can speculate about the ways in which those capabilities could exacerbate cyber risks at a global scale. For example:

- **Orchestration of simple attacks at scale.** AI could quickly and autonomously generate the infrastructure and execute on known attacks at a scale and speed that was previously infeasible. This could allow vulnerabilities to be exploited faster than patches can be applied.
- **Identifying new vulnerabilities.** Google's "Big Sleep" AI model recently located a brand-new software vulnerability that human auditors had missed.[41] Fortunately, they used this discovery to patch the vulnerability. However, if attackers can gain an upper hand in automated vulnerability discovery, this could render existing systems far less secure.
- **Identifying new classes of vulnerabilities.** Future models may unearth subtler flaws, e.g., new side-channel attacks such as Van Eck phreaking (listening for electromagnetic emissions) or acoustic cryptanalysis (e.g. inferring keyboard strokes from audio)—creating threats defenders have never faced.[42],[43]
- **Malware with "AI copilots".** Malware authors can already use AI models for help rewriting code or crafting phishing emails. As AI models become more capable and efficient, it may be possible to ship malware with a sophisticated onboard AI agent able to probe, adapt, and move laterally without supervision.
- **Increasingly autonomous malware scenarios.** With on-board AI agents, increasingly powerful autonomous attacks may be possible. An agent might infiltrate a government network, map its org-chart from email access, use surreptitious recordings to fabricate deep-fake zoom calls of officials to escalate privileges, and silently exfiltrate sensitive data—all without human cues. Though hypothetical, AI models are already capable of several parts of this chain today.
- **Speculative scenarios of fully autonomous models pursuing their own goals.** The scenarios above implicitly assume that the AI models are ultimately acting under some degree of human direction and control. In the worst-case scenario, AI models with poorly specified goals may start to pursue instrumentally valuable[44] goals such as self-preservation and resource acquisition, and use sophisticated cyber capabilities to achieve these ends. This could include, for example, exfiltrating their own weights and hacking to acquire resources.

Balancing speculation with evidence

The largest scale risks may lie at the interaction of the speculative AI capabilities and targets described above. However, it is important to emphasize that most extreme forecasts remain uncertain, and based on AI capabilities that are well beyond those of current models. Even granting such capabilities, it is still important to consider the offense-defense balance. Cyber offense may gain speed and capability, but the same models are also likely to support defenders in rapid detection, patch generation, and automated incident response.[45] Assessing both evolving model capabilities and the offense-defense balance under future capability curves is therefore essential. Finally, the above discussion has made no reference to the intentions and motivations of attackers, and considering who would want to execute particular attacks is also an important consideration.

41 Google. (2024, November 1). Project Zero: From Naptime to Big Sleep: Using Large Language Models To Catch Vulnerabilities In Real-World Code. Project Zero. https://googleprojectzero.blogspot.com/2024/10/from-naptime-to-big-sleep.html

42 Van Eck phreaking definition – Glossary | NordVPN. (2023, August 28). https://nordvpn.com/cybersecurity/glossary/van-eck-phreaking/

43 What Is Acoustic Cryptanalysis? - ITU Online IT Training. (2024, March 27). https://www.ituonline.com/tech-definitions/what-is-acoustic-cryptanalysis/

44 Tsvi Benson-Tilsen & Nate Soares. (2016). Formalizing Convergent Instrumental Goals (The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence AI, Ethics, and Society: Technical Report WS-16-02). https://cdn.aaai.org/ocs/ws/ws0218/12634-57409-1-PB.pdf

45 Seizing AI's Trillion Dollar Cyber Opportunity. (2025, July 9). Cato Institute. https://www.cato.org/blog/seizing-ais-trillion-dollar-cyber-opportunity

AI Models as a New Systemic Attack Surface

The above discussion has focused on AI's use as a tool or agent in cyber offense and defense. As AI is increasingly embedded in critical workflows, vulnerabilities in the models themselves could become vectors of large-scale risk. For example:

- **Adversarial misuse.** Prompt injection and jailbreaks already make public chatbots reveal restricted content and engage in unexpected behaviors. In safety-critical settings—e.g., an AI scheduling power-plant maintenance—such manipulation could cause real damage.
- **Data-poisoning and supply-chain attacks.** By incorporating "poisoned" data into the training corpus of AI models, attackers can manipulate the outputs of models. AI developers often may not carefully scan their training data for adversarial examples, leaving them vulnerable to contaminated data.
- **Hidden back-doors.** Anthropic's sleeper-agent work (cited above) shows how deceptive behaviors can potentially be embedded into models. If such a back-doored model was widely used, it could make a large number of users vulnerable to exploitation.

In all of the above cases, the risk is higher when critical systems rest on a small number of AI models. Today, a handful of foundation models increasingly underpin an exploding ecosystem of downstream applications. Such an "IT monoculture"[46] means that an exploit found in a single popular model is likely to have a very wide blast radius.

Potential Policy Responses

Policy does not move at the pace of technology, and making policy in the face of a rapidly evolving technology is especially challenging. Some potential policy responses to address the largest-scale risks include:

- **Scale government AI capacity.** The recently renamed NIST Center for AI Standards and Innovation (CAISI) is tasked with evaluating commercial models for security and misuse. Expanding CAISI's compute and hiring authority would let it probe frontier systems in greater depth.
- **Invest in defensive AI R&D.** The U.S. Department of Energy is funding AI tools that monitor grid telemetry, simulate attack chains, and recommend mitigations in real time.[47] Increasing support for such programs across the government could move the offense-defense balance in a favorable direction.
- **Promote or mandate pre-deployment model evaluations and disclosure.** Frontier developers already conduct red-team tests voluntarily; regulation could formalize this, requiring the companies, or independent auditors, to evaluate the cyber risks presented by new models.
- **Promote diversity and redundancy in critical AI stacks.** To blunt monoculture risk, governments could incentivize multiple vendors, open standards, and "hot-swap" contingency plans so that failure or compromise of one model does not cascade.
- **Build AI-cyber crisis-response playbooks.** National cyber exercises could include scenarios where an AI agent drives a cyber attack or where a critical model shows anomalous behavior.

Questions for Discussion

- **Future capability curve:** Which cyber-relevant AI abilities are most likely to mature in the next five years?
- **Degree vs. kind:** Will AI transform cyber risk incrementally or introduce fundamentally new threat classes?
- **Offense–defense balance:** Under what conditions could defensive AI outweigh offensive gains and vice versa?
- **Highest-impact sectors:** Which areas—nuclear, power grid, healthcare, information ecosystem—could be most significantly impacted by AI-cyber capabilities?
- **Speculative vs. plausible:** Which of the outlined extreme scenarios do you find most and least credible, and why?

46 Risks Associated with IT Monoculture Needs Further Examination. (n.d.). Retrieved July 14, 2025, from https://www.centerforcybersecuritypolicy.org/insights-and-research/risks-associated-with-it-monoculture-needs-further-examination

47 Oakland, S. (2025, June 5). DOE Accelerates AI Research to Defend Critical Infrastructure. GovCIO Media & Research. https://govciomedia.com/doe-accelerates-ai-research-to-defend-critical-infrastructure/

- **Framework adequacy:** Do current evaluation methods capture the riskiest AI-cyber interactions, or do we need new metrics to understand emerging capabilities?
- **Research priorities:** What technical or policy research would most reduce uncertainty about large-scale AI-cyber risks?
- **Governance levers:** How should policy levers evolve if frontier models start enabling dangerous cyber autonomy?

PROTECTING AI MODEL WEIGHTS IN A HIGH-STAKES ERA

HAMZA CHAUDRY, FUTURE OF LIFE INSTITUTE

Frontier AI model weights—the numerical parameters that encapsulate a system's intelligence—are fast becoming strategic national-security assets. U.S. officials already view the most advanced models as potential sources of decisive diplomatic, technological, and economic advantage. If adversaries obtain these weights, they gain unrestricted access to the model's full capabilities for sabotage or other harm.

Training such systems now costs hundreds of millions of dollars—and may soon cost billions. The expense of stealing an existing model is orders of magnitude lower, making these weights an especially attractive target for espionage and cyber-theft. Panel 3, therefore, focuses on the methods of securing frontier models against unauthorized access, theft, and misuse.

Risks and Threat Vectors

A recent RAND study, titled *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*, identified 38 distinct attack vectors by which determined attackers might attempt to steal or compromise AI model weights. These range from relatively mundane methods to highly sophisticated schemes.[48] Key examples include:

- **Insider Threats.** An employee of an AI lab could exfiltrate the model (e.g. downloading weights onto a drive). This could be motivated by financial gain or coercion.
- **External Cyberattacks.** Hackers might breach the company's network or cloud infrastructure to steal model files. This could be a criminal group or a state-sponsored advanced persistent threat.
- **Supply Chain and Technical Attacks.** Attackers might target vulnerabilities in the machine learning pipeline—e.g. intercepting model files in transit, exploiting bugs in version control or API interfaces to extract the model, or using adversarial queries to reconstruct the model (model extraction attacks).
- **Physical Breaches and Tampering.** In extreme cases, someone could attempt a physical intrusion (breaking into data centers or seizing hardware where models are stored).[49]

Further, not all attacks are equal in terms of capability and motivation. For example, the RAND study outlines five tiers of attacker sophistication, ranging from low-budget amateurs to cybercriminal gangs and top-tier nation-state intelligence services. Less capable actors might only attempt simple social engineering or exploit publicly known bugs, whereas a nation-state could deploy custom hacking tools, insider recruitment, or supply chain compromises. There are several motivations that may push attackers to acquire frontier model weights, including economic espionage, national security advantage, and the desire to do reputational damage to a given model developer or provider.[50]

If a frontier model's weights are stolen by a rival or released publicly without safeguards, several risks emerge:

- A hostile nation could leverage the model to boost its own AI capabilities (narrowing the U.S. lead in AI) or integrate it into military and espionage applications. The National Security Agency (NSA) has raised alarms about "foreign adversaries' interest in targeting U.S. AI models" via espionage. NSA officials believe we may face a "swarm of nation-state espionage" attempts against American AI companies and research centers.
- Criminal or terrorist groups obtaining a cutting-edge model could enable a range of malicious activities. The model could be fine-tuned for disinformation or cybercrime, for example, without any of the safety filters the original developers put in place.
- The *proliferation* of powerful models raises longer-term concerns as well. If many actors have access, it becomes harder to prevent misuse. It might also accelerate an international AI arms race, where countries

48 Nevo, S., Lahav, D., Karpur, A., Bar-On, Y., Bradley, H. A., & Alstott, J. (2024). Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models. https://www.rand.org/pubs/research_reports/RRA2849-1.html
49 Ibid.
50 Ibid.

or groups feel pressured to build or deploy advanced AI quickly (possibly without adequate safety testing) because others have the capability.[51]

Hardware-Level Security Measures

Modern hardware solutions can form the foundation of responding to the threats outlined above and protecting AI model weights. One important approach is confidential computing—leveraging hardware-based Trusted Execution Environments (TEEs) to keep data (including model weights) encrypted in memory and only decrypted within a secure enclave during computation. These technologies ensure that even if an attacker compromises the operating system, they cannot read the model weights in plaintext form. Major cloud providers are already adopting this: for example, Google Cloud offers Confidential Virtual Machines.

Beyond TEEs, hardware "roots of trust" play a critical role. Secure boot processes, cryptographic chip identifiers, and attestation mechanisms ensure that only trusted, untampered firmware and software can load a model. This prevents low-level malware from subverting the environment to dump or alter weights. Many AI data center systems now incorporate hardware security modules and TPMs (Trusted Platform Modules) to manage cryptographic keys for encryption of data at rest and in transit, and to verify system integrity at boot.

Another avenue is designing secure AI accelerators and modules explicitly for sensitive workloads. This includes R&D into tamper-resistant hardware, including chips and server boards designed to detect and resist physical tampering. For instance, there are proposals for GPU boards with built-in sensors that erase memory if the case is opened without authorization.

It is important to note that confidential computing at scale is still maturing. Confidential GPU clouds exist at the rack level, but extending encryption-in-use across entire multi-GPU clusters (for training trillion-parameter models) remains a challenge. Nonetheless, incorporating hardware security now—even if partially—reduces the attack surface significantly. Hardware security provides a critical backstop: even if adversaries penetrate networks or have a mole on the inside, strong hardware protections can make extracting the actual model weights far more difficult.

Software Controls and Access Management

Software-level controls are equally important in safeguarding model weights. These include how we store, encrypt, manage access to, and monitor the use of models in software systems. Examples include:

- **Encryption and Key Management.** AI developers can encrypt model weights both while stored and while being transmitted. Further, they can store the encryption keys in hardware security modules or secure vaults and rotate them regularly. Decryption could even only be allowed inside protected memory that is accessible to authorized systems or personnel. Finally, proper key management policies are crucial so that only designated systems or personnel can initiate model decryption.
- **Strict Access Controls.** AI developers can apply role-based access control so that only a very small set of services or users can read or modify the weights. Keeping the weights in encrypted repositories that are reachable only through tightly controlled application programming interfaces, and never through direct file downloads, provides additional access controls.
- **Authentication and Authorization.** AI developers can create a series of actions, including enforcing strong multi-factor authentication or hardware security keys, granting access only for the time and context needed, and requiring a second approval for sensitive actions such as exporting a model to a new environment.
- **Logging and Monitoring.** AI developers can record every access or change to model weights in tamper-proof logs, analyze those records in real time, and trigger immediate alerts when unusual behavior—such as large transfers at odd hours or repeated failures—occurs.
- **Versioning and Integrity Checks.** AI developers can maintain strict version controls for the AI model weights, allowing unauthorized changes to be detected. A check for any changes should be run each time the model

51 Ibid.

loads. AI developers should also keep append-only logs and secure backups, and roll back or retrain the model if tampering is detected (although note this could be expensive).[52]

Infrastructure and Data Center-Level Defenses

The physical and network infrastructure hosting frontier models is the next line of defense. Given the scale of advanced AI training (often clusters of specialized hardware in data centers), special architectural measures are needed, which could include:

- **Network Segmentation & Isolation.** House training clusters on sealed networks with tight internal firewalls. There should be no direct internet path for an attacker to extract models, and minimal lateral traffic.
- **Data-Center Physical Security.** Guard facilities housing AI infrastructure with multi-layer badges, cameras, and patrols to block unauthorized entry or disk removal.
- **Defense Against Physical Tampering.** Within the data center, additional measures can protect against an attacker who gains brief physical access.
- **Redundancy & Resilience.** Keep encrypted weight backups offline/off-site and maintain robust power, fire, and disaster recovery plans so sabotage or outages don't wipe models.

Mitigating Insider Threats

Insider threats merit special focus, because trusted insiders can bypass many external-facing controls. Mitigation starts from hiring and extends through daily operations, potentially including:

- **Personnel Vetting & Reliability Programs.** Run thorough background/reference checks (and clearances for ultra-sensitive roles) and keep re-evaluating for risk signals.
- **Need-to-Know Access Limitation.** Give model-weight or code access only to those who truly need the specific slice they're working on.
- **Two-Person Rule and Dual Controls.** Make any high-impact action (e.g., exporting weights, pushing to prod) require a second approver or joint action.
- **Monitoring & Behavioral Analytics.** Track anomalous system use, big data moves, privilege jumps, plus soft signs of disgruntlement; verify every action.
- **Audit Trails & Regular Audits.** Log all privileged activity and review logs, permissions, and dormant accounts routinely—with outside spot-checks when possible.

Policy Responses

In 2024, the White House released a National Security Memorandum (NSM) on AI—the first comprehensive U.S. national security strategy for AI.[53] The NSM and an accompanying framework call on the Department of Defense and Intelligence Community to integrate AI security into their operations—ensuring the DoD can securely adopt AI, red-team its systems, and guard against adversary AI.[54]

Another pillar of U.S. strategy has been to prevent adversarial nations from acquiring the most advanced AI capabilities, through export controls and allied coordination. Starting in 2022, the U.S. Commerce Department (Bureau of Industry and Security) imposed export controls on high-end semiconductor chips crucial for training large AI models (such as certain NVIDIA GPUs), specifically restricting exports to China and other strategic rivals.[55]

52 Ibid.
53 The White House. (2024, April 30). National Security Memorandum on Critical Infrastructure Security and Resilience. The White House. https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2024/04/30/national-security-memorandum-on-critical-infrastructure-security-and-resilience/
54 The White House. (2024). Framework to Advance AI Governance and Risk Management in National Security. https://ai.gov/wp-content/uploads/2024/10/NSM-Framework-to-Advance-AI-Governance-and-Risk-Management-in-National-Security.pdf
55 Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification. (2022, October 13). Federal Register. https://www.federalregister.gov/documents/2022/10/13/2022-21658/implementation-of-additional-export-controls-certain-advanced-computing-and-semiconductor

These controls were tightened in 2023 to cover more chip types and to limit Chinese access to cloud-computing resources for AI. Such 'compute security' restrictions may also help enhance model weight protection.[56]

Policymakers must start treating frontier AI models as a sensitive technology to be actively safeguarded. This includes securing the AI supply chain (chips and models), hardening AI systems against attacks, and ensuring rigorous evaluation of advanced AI before it is deployed. The challenge will be balancing security with innovation—guarding against threats without unduly hampering beneficial AI development. So far, the approach leans toward collaboration with industry, backed by the threat of regulation or enforcement if needed.

Conclusion

Frontier AI models represent a new frontier for security policy. We must treat AI model weight protection as a central pillar of AI governance and national security strategy, not an afterthought. This means investing in advanced security R&D (for example, developing tamper-proof AI hardware and new encryption techniques for models), updating policies and possibly laws (to ensure companies have both the obligation and support to protect models), and fostering international cooperation to prevent a global "AI Wild West" where stolen models fuel proliferation.

Questions for Discussion

- How do we incentivize AI labs and cloud providers to quickly adopt measures like confidential computing, stringent access limits, and continuous monitoring?
- What are the key bottlenecks to implementing various security measures discussed above, and what role could the government play in alleviating these bottlenecks?
- Should any of these best practices be made mandatory for frontier model developers (through regulation or procurement requirements) if voluntary uptake is slow?
- What are the trade-offs inherent in increased security measures, and how should we balance them? For example, how should we balance mitigating insider threats with maintaining a diverse and inclusive development ecosystem, often supercharged by foreign talent?
- As frontier models grow more powerful (with training runs using ever more data and compute), we may approach scenarios where AI systems exhibit qualitatively new abilities (problem-solving, autonomy, etc.). How should AI security requirements for models scale as capabilities evolve? Are there particular thresholds where we should increase security requirements?

56 Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections. (2023, October 25). Federal Register. https://www.federalregister.gov/documents/2023/10/25/2023-23055/implementation-of-additional-export-controls-certain-advanced-computing-items-supercomputer-and

Interested to learn more about our AIxGlobal Risk Nexus Series?

Visit FAS.org to learn more about our upcoming events, publications and Global Summit 2026.

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.**