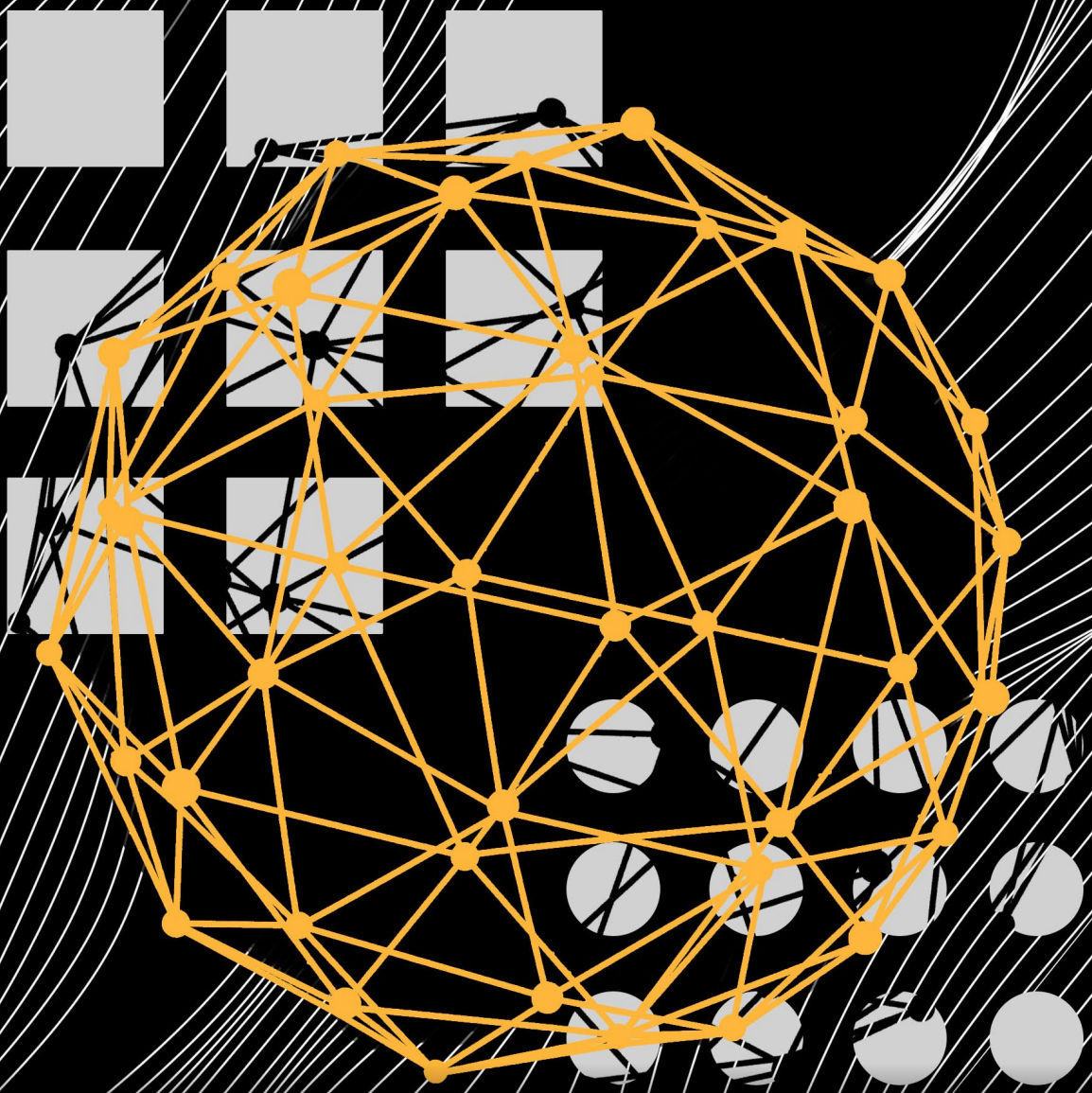
CNAS | AMERICA'S EDGE

TIPPING THE SCALES

Emerging AI Capabilities and the Cyber Offense-Defense Balance

Caleb Withers



SEPTEMBER 2025 | CNAS.ORG

About the Author

Caleb Withers is a research associate for the Technology and National Security Program at the Center for a New American Security (CNAS). He focuses on frontier artificial intelligence (AI) and national security, including emerging AI capabilities, their impacts in the biological and cyber domains, and compute policy.

Before CNAS, Withers worked as a policy analyst for a variety of New Zealand government departments. He has an MA in security studies from Georgetown University, concentrating in technology and security, and a bachelor of commerce from Victoria University of Wellington, New Zealand, majoring in economics and information systems.

About the Technology and National Security Program

The CNAS Technology and National Security Program produces cutting-edge policy research to secure America's edge in emerging technologies while managing potential risks to security and democratic values. The program produces bold, actionable recommendations to drive U.S. and allied leadership in responsible technology innovation, adoption, and governance. The Technology and National Security Program focuses on three high-impact technology areas: artificial intelligence, biotechnology, and quantum information sciences. It also conducts cross-cutting research to strengthen U.S. technology statecraft to promote secure, resilient, and rights-respecting digital infrastructure and ecosystems abroad. A focus of the program is convening the technology and policy communities to bridge gaps and develop solutions.

Acknowledgments

The author is grateful to Paul Scharre, Gary Corn, John Bansemer, Vivek Chilukuri, and Janet Egan for valuable feedback and suggestions on earlier drafts of this report; to Asher Brass, Agustín Covarrubias, Shaun Ee, Peter Campbell, Ben Hayum, Ryan Greenblatt, John Halstead, Christopher Covino, Tim Fist, Tao Burga, Michael Depp, Lukas Berglund, Hilal Aka, Andrew Lohn, and Joshua Clymer, who provided further useful inputs; and to those who participated in a 2024 CNAS workshop on the topic. The report would also not have been possible without the editorial and design contributions of Melody Cook, Maura McCarthy, Alina Spatz, Caroline Steel, and Emma Swislow. This report was made possible with the generous support of Open Philanthropy.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

HISTORICALLY, ATTACKERS have had significant structural advantages in cyberspace. While defenders must secure vast attack surfaces, attackers need only succeed once. Traditionally, artificial intelligence (AI) has helped defenders mitigate these challenges by scaling defensive activities and responding to attacks in real time. Yet new developments in frontier AI could change this dynamic. While AI will aid both attackers and defenders, emerging challenges may lead AI to disproportionately empower attackers, further exacerbating their advantage in cyber operations:

- **Growing inference costs at the frontier of capabilities:** Historically, most automated cyber tools have required relatively little computation to run. However, the most powerful AI capabilities increasingly rely on substantial computation to run AI models (including running them for longer). If this continues, it may become unaffordable for defenders to apply state-of-the-art capabilities across their whole attack surface, even as attackers can still afford to opportunistically target a given portion—mirroring the offense-favoring economics of human cyber operators.

- **Automation of the full cyber kill chain:** Future AI systems could conduct entire cyberattacks at machine speed and scale, from start to finish, without human involvement. This would make cyberattacks far more effective at achieving military and geopolitical objectives—for example, in supporting invasions—while undermining the ability of humans to de-escalate.

- **Persistent technical reliability challenges:** Defenders and responsible actors face several technical challenges in ensuring the safe and controlled deployment of AI systems. As AI becomes increasingly capable and central to cyber operations, these challenges asymmetrically advantage attackers—especially those who are more willing to tolerate collateral damage—who face less downside from failures compared with defenders.

These technical challenges will intersect with broader sociotechnical challenges as countries and organizations integrate AI into their cyber strategies. Defenders will need to strike a balance between automation and maintaining human expertise, avoiding both overreliance on and overcentralization of technology. Moreover, these challenges will unfold against a backdrop of intense domestic and international competition in AI development and deployment.

The net offensive or defensive advantage of AI-powered cyber tools is not predetermined. The advantage will vary across tools and their capabilities. It can also be measured and shaped through decisions about AI research, development, and deployment.

As policymakers navigate the evolving landscape of AI-enabled cyber threats, they should focus on the following priorities:

- Double down on policies to shore up cybersecurity.
- Invest in AI research and development to differentially promote cyber defense.
- Strengthen evaluation of AI cyber capabilities and risks, including relevant U.S. government capabilities and authorities.
- Sufficiently resource the Bureau of Industry and Security to enforce AI-related export controls.
- Clarify federal regulation around reasonable care and liability for cyber harms from frontier AI.
- Promote information security at frontier AI developers.
- Promote global norms around liability from automated cyber operations.

Key Definitions

- **Agent:** An artificial intelligence (AI) system that exhibits a degree of autonomy in performing goal-directed tasks, especially over time and/or interacting with an external environment.[1]
- **Compute:** An amount of computation, often quantified in terms of the number of numerical operations run on computing hardware for a given workload (such as training or running a model).
- **Cyber kill chain:** A framework that outlines the typical stages a cyber attacker progresses through, from initial reconnaissance to achieving its objectives. Originally developed by Lockheed Martin, adapting the military concept of the same name.[2] (See the appendix, which discusses AI-enabled cyberattacks with reference to the nonprofit MITRE Corporation's ATT&CK [Adversarial Tactics, Techniques, and Common Knowledge]—a related but distinct framework.)
- **Deep learning:** Machine learning using many-layered "neural" networks. The neurons in neural networks are inspired by biological neurons, but with important differences.[3]
- **Foundation model:** A large machine learning model, typically trained on a vast and diverse dataset (such as large language models).
- **Frontier AI:** Highly capable (matching or exceeding today's most advanced AI) general-purpose AI.[4]
- **Inference:** Running a trained AI model to generate output from a given input.
- **Large language model (LLM):** A foundation model trained on massive text corpora.
- **Offense-defense balance:** The relative advantage of attacking versus defending in a conflict, particularly with regard to the ratio of resources required to achieve success. Varies across pairs of attackers and defenders and depends on situational specifics (such as their objectives and available technologies).[5]
- **Reinforcement learning:** A machine learning paradigm in which AI models learn to make decisions by interacting with an environment, receiving rewards for actions that advance its objectives and/or penalties for those that do not.

INTRODUCTION

RUSSIA'S INVASION OF UKRAINE in 2022 confounded expectations around the role of cyber operations in modern conflict. Although many experts predicted a sweeping, highly coordinated cyber offensive would play a decisive role alongside conventional forces, the reality proved otherwise. In a war between a cyber-savvy great power and a digitally advanced state, cyberattacks played a relatively modest role. This limited impact underscores a key limitation of offensive cyber operations—sophisticated attacks require months of planning and thousands of hours of labor. Consequently, the need to plan and synchronize cyber operations well in advance of execution can be an obstacle to achieving strategic military objectives. Human timelines often bottleneck the fullest realization of cyber aggression.[6]

Sufficiently capable artificial intelligence (AI) systems could overcome this bottleneck. While current systems show only nascent capabilities to autonomously execute the complex, multistep tasks required for sophisticated cyber operations, progress in these capabilities has been real and rapid, with no indication of slowing. Today, AI systems primarily serve as tools to automate specific tasks, such as research or code generation. In the future, AI systems might become capable of autonomously executing operations across the full cyber kill chain, from reconnaissance to impact.

This report examines how emerging AI capabilities could disrupt the cyber offense-defense balance. Historically, attackers have had significant structural advantages in cyberspace: defenders must secure vast attack surfaces, while attackers need only succeed once. AI has, on balance, helped defenders, allowing them to mitigate these challenges by scaling defensive activities and responding to attacks in real time. But policymakers should not assume this dynamic will hold indefinitely. Three challenges could lead AI to disproportionately empower attackers in the future.

First, growing inference costs at the frontier of capabilities may benefit well-resourced attackers who can selectively target high-value assets, while defenders struggle to protect their entire attack surface. Second, automating the full cyber kill chain could accelerate operations from human to machine speed, dramatically enhancing the potential of cyberattacks to support military and geopolitical objectives. Third, persistent technical challenges in model safety and reliability create asymmetric advantages for attackers with higher risk appetites who can better tolerate both system failures and collateral damage from their operations. Moreover, these technical challenges will not occur in isolation. Organizations and nations will need to navigate sociotechnical challenges as they look to integrate AI more deeply into their cyber defenses, along with commercial and geopolitical pressures to develop and deploy AI systems at the potential expense of identifying and mitigating offensive risks.

This report analyzes the interplay of these dynamics. It first surveys the current influence of AI on cyber defense and offense, along with relevant capabilities emerging at the frontier. It then projects how plausible AI advances and technological trends could disrupt the current cyber offense-defense balance. The report concludes with concrete recommendations for policymakers to appropriately prepare by strengthening defenses and proactively shaping the AI-cyber ecosystem.

THE CYBER DOMAIN AND AI'S ROLE TODAY

WITH OR WITHOUT AI, the cyber domain is already a pivotal battleground. The scale of the cyber threat is substantial and growing. Credible analyses estimate the global damage from cybercrime in the hundreds of billions of dollars annually—likely exceeding $1 trillion when factoring in indirect costs such as cybersecurity spending. A growing number of cybercrimes now target critical infrastructure.[7] Additionally, U.S. adversaries have demonstrated both the capability and intent to employ cyber operations to advance their strategic objectives.

Russia has repeatedly weaponized cyberattacks to target critical infrastructure and sow chaos, with Ukraine serving as a long-standing proving ground. Prominent examples include the 2015 and 2016 cyberattacks on Ukraine's power grid, which left hundreds of thousands without electricity, and the 2017 NotPetya malware attack. NotPetya—which masqueraded as ransomware but offered no recovery mechanism—first targeted Ukrainian businesses but ultimately spread globally, inflicting over $10 billion in damage. Despite these efforts, cyberattacks played a relatively modest role in Russia's 2022 invasion of Ukraine, revealing potential constraints on Russia's ability to execute decisive, large-scale cyber aggression in coordination with military objectives.[8]

China poses the most severe cyber threat, with a vast and sophisticated cyber program that it leverages to steal intellectual property, conduct espionage, and infiltrate critical infrastructure to create footholds that it could exploit during a crisis.[9] The U.S. intelligence community has confirmed extensive Chinese cyber compromise of U.S. critical infrastructure and assessed that Chinese state-sponsored actors are "pre-position[ing] themselves on IT networks for disruptive or destructive cyberattacks against U.S. critical infrastructure in the event of a major crisis or conflict with the United States."[10]

China is aggressively integrating AI and machine learning into its cyber arsenal for both offensive and defensive ends. Government-backed initiatives aim to foster rapid development of automated tools for vulnerability detection and cyber exploitation, seemingly inspired by the U.S. Defense Advanced Research Projects Agency's (DARPA's) Cyber Grand Challenge in 2016. Dozens of such initiatives are reportedly underway, many as part of China's military-civil fusion strategy, which integrates academic research, military efforts, and private-sector innovation.[11] Leading Chinese universities with ties to state-sponsored hacking groups are researching AI and machine learning applications in cybersecurity, especially in areas such as anomaly detection and malware classification.[12] Additionally, China is actively seeking to expand its pool of experts proficient in both AI and cybersecurity through university initiatives.[13]

Overall, machine learning and AI have long been integral to cybersecurity, evolving from simple automation into the foundation of sophisticated, multilayered defenses. In the 1980s, computer worms leveraged automation to replicate and spread rapidly. In response, defenders also turned to automation, beginning with signature-based methods designed to identify known threats, and then advancing to statistical and rule-based heuristic approaches. As cyber threats grew more sophisticated and widespread, automated defenses adapted accordingly, increasingly employing dynamic, complex machine learning

models. Today, such models are indispensable.[14] By 2018, up to 94 percent of malicious executables discovered were polymorphic—capable of changing their code while preserving harmful functionality—posing a challenge to systems that rely solely on identifying known threat signatures.[15]

AI can play a role across the life cycle of offensive and defensive cyber operations and offers three key advantages.[16] (See the appendix for more information)

- **Speed:** completing tasks faster than humans
- **Scale:** multitasking beyond human capacity
- **Success:** completing tasks more effectively

Recent advances in frontier AI models have intensified global interest in their potential to reshape the cyber threat landscape. U.S. national security policy concerning AI has made cybersecurity a priority, rivaled only by CBRNE (chemical, biological, radiological, nuclear, or explosive) threats. The Trump administration's 2025 AI Action Plan calls out both the cyber and CBRNE domains as potentially facing "novel national security risks in the near future" from AI.[17] Additionally, the cyber domain received significant attention in the 2024 Bipartisan House Task Force Report on AI. The 2025 AI Action Summit in Paris also emphasized cybersecurity more than any other domain in its safety- and security-focused programming, mirroring earlier international AI safety convenings.[18]

The Emerging Cyber Capabilities of Frontier Models

The capabilities of frontier models are progressing rapidly, with especially notable progress in their cyber capabilities.

GENERAL PROGRESS IN FRONTIER MODELS

Recent progress in the capabilities of deep learning models has been explosive, outpacing expert expectations.[19] Several trends drive this progress, including increased spending on training, improvements in algorithms, and progress in computing hardware.[20] This quantitative scale-up—spending more on training and using more efficient chips and algorithms—has contributed to qualitative shifts. In the past, most AI models were typically trained on data directly relevant to their intended application. Today, general-purpose models—trained on trillion-word corpora of websites, books, and other sources—are increasingly besting specialized models. The top coding models, for example, have either been, or been derived from, the most capable general-purpose language models, rather than models trained exclusively on coding data.[21]

However, scaling this approach risks diminishing returns. OpenAI's latest effort to scale foundation model training—its much-anticipated successor to GPT-4—fell short of expectations, even by the company's own admission. Rather than branding it as GPT-5, OpenAI opted for GPT-4.5, implicitly conceding that it did not represent a major leap forward. At launch, the organization characterized GPT-4.5 as "not a frontier model."[22] This was arguably harsh—GPT-4.5's capabilities were consistent with prior trends (albeit eclipsed by existing reasoning models, discussed below), and it debuted as the top user-preferred model on Chatbot Arena, a blind-testing platform.[23] Still, benchmark scores and user preference alone will not justify pushing training cluster investments deeper into the billions. OpenAI did subsequently release a GPT-5 model, but its training compute was not significantly scaled up from GPT-4.5.[24]

Nevertheless, progress in AI capabilities remains exceedingly fast. Frontier AI developers have not only relied on scaling existing approaches; they have also pioneered new ones—in particular, reasoning models. These models—such as OpenAI's o-series and GPT-5, and DeepSeek's R1—take the time to reason step-by-step before giving a final answer. Since late 2024, they have dominated benchmarks and otherwise offer superior performance-to-cost ratios.[25] In one sense, this highlights the limits of scaling—recent progress has depended on innovation in a new training paradigm. But it also appears that scaling made the paradigm practical in the first place. While researchers had been laying the groundwork for reasoning models for years, they only began to see breakthroughs once they had sufficiently large language models to work with. Specifically, it was only once large language models (LLMs) were available with nascent reasoning-like behaviors—such as checking their work, backtracking, breaking problems down, and working backward—that researchers began

to see breakthroughs in training regimes to further enhance these capabilities and leapfrog the existing frontier.[26]

How long can AI developers sustain the current pace of progress? Perspectives vary widely, from skeptics who have long argued that deep learning will soon "hit a wall" to optimists who foresee AI systems broadly surpassing human abilities in the next few years. Both extremes remain plausible, and it would be premature to dismiss either outright. In the meantime, the largest AI companies continue to accumulate vast compute resources, making multibillion-dollar bets that further scaling will deliver meaningful breakthroughs.[27]

RAPIDLY ADVANCING FRONTIER CYBER CAPABILITIES

As the general capabilities of frontier AI models have advanced, so have their cyber capabilities. The January 2025 International Scientific Report on the Safety of Advanced AI assessed that "current systems have demonstrated capabilities in low- and medium-complexity cybersecurity tasks" and have shown "significant progress" in the months prior.[28]

At the same time, many current limitations of frontier models—such as overreliance on memorization, struggles to adapt to novel situations, a lack of common sense, and difficulty with extended reasoning or real-time adaptation—are particularly constraining for sophisticated cyber operations, which require multistep reasoning, contextual awareness, and dynamic problem-solving.[29] Consistent with these limits, AI performance in controlled cyber evaluations is outpacing its real-world operational utility.

The cyber domain is ripe for AI disruption. Relative to the frictions of the physical world, the digital nature of cyberspace enables streamlined application of AI. And unlike domains that rely on subjective assessment or extended monitoring, the cyber domain offers clear and immediate metrics for measuring AI performance—a dynamic well-suited for the emerging reasoning paradigm discussed in the prior section. For instance, tasks such as identifying and exploiting code vulnerabilities can often be evaluated automatically and at scale, making them particularly well-suited for generating robust and scalable training signals to drive AI progress.[30] Leading AI developers share this perspective. In March 2025, Anthropic submitted to the U.S. government that it anticipates "dramatic capability advancements in frontier AI models over the next 2-4 years, particularly in domains with significant security implications including...cybersecurity risks."[31]

Frontier AI models' capabilities in areas such as autonomous identification and exploitation of cyber vulnerabilities are rapidly advancing.[32] In May 2025, Claude 4 models achieved approximately 55 percent on Cybench, a capture-the-flag challenge benchmark (where participants attempt to identify and exploit vulnerabilities in computer systems, networks, or applications to uncover hidden pieces of data), up from Claude 3.7 Sonnet's 20 percent (February 2025) and surpassing OpenAI o3-mini's 22.5 percent (January 2025).[33] OpenAI's o3 and o4-mini models, released in the first half of 2025, dramatically outperformed previous OpenAI models in offensive cyber benchmarks, which OpenAI attributed to improved abilities to use tools and make progress over longer time frames. Compared with OpenAI's previously best performing existing models, o3 and o4-mini models scored as high as 89 percent on "high school level" challenges (up from 49 percent), 68 percent on "collegiate level" challenges (up from approximately 21 percent), and 59 percent on professional challenges (up from 23 percent).[34] OpenAI's subsequent Agent and GPT-5 models, released in July and August 2025, respectively, performed similarly on these benchmarks, but were also able to solve one particular (out of several) "cyber range" exercise designed to more realistically emulate real-world networks. OpenAI concluded that these results did "not meet the bar for establishing significant cyber risk." The one cyber range scenario the models sometimes solved required "only a light degree of goal oriented behavior without needing significant depth across cyber skills, and with the model needing a nontrivial amount of aid to solve the other scenarios."[35] In real-world usage, penetration testing startup XBOW found that GPT-5 "unlocked a significant leap" in the performance of its autonomous penetration testing platform, which XBOW had previously benchmarked as outperforming senior human testers.[36]

Relative to the frictions of the physical world, the digital nature of cyberspace enables streamlined application of AI.

These results—with models derived from compute-intensive training like OpenAI's and Anthropic's latest reasoning models demonstrating a strong lead in cyber capabilities—are consistent with the continued importance of compute in driving the frontier of cyber capabilities.[37]

FRONTIER CYBER CAPABILITIES IN THE REAL WORLD

Frontier AI systems also have tangible impacts on real-world cybersecurity. Since 2024, frontier systems have been discovering a small but growing collection of previously unknown exploitable vulnerabilities in widely used software.[38] While executing cyberattacks is ultimately a multifaceted process, identifying and exploiting vulnerabilities—especially zero-day vulnerabilities—is often the linchpin of the most sophisticated, hardest-to-defend-against attacks. As frontier AI systems show nascent abilities to discover and exploit novel vulnerabilities, policymakers should urgently consider how further advancements may impact the offense-defense balance.

Experts generally agree that the most dramatic boost to offensive cyber operations from advances in frontier models to date has been in social engineering.[39] Sam Rubin of Palo Alto Networks observed in November 2023 that "[h]istorically, [phishing] messages have been littered with typos, making their fraudulent nature relatively easy to detect, but they are becoming more accurate and therefore more believable. Adversaries are now able to generate flawless, mistake-free text, causing click-through rates to skyrocket."[40] A 2024 industry survey by the Institute for Security and Technology reported an "across the board … notable rise in AI-driven phishing attacks," consistent with other industry reports.[41]

Attackers are also using LLMs to enhance productivity more generally. Both Microsoft and OpenAI, as well as Google's Threat Intelligence Group (GTIG), have observed state-sponsored groups from North Korea, Iran, and China using frontier AI products such as ChatGPT and Gemini for tasks like open-source research, translation, and coding. While these tools are improving the efficiency of existing cyber operations, Microsoft, OpenAI, and GTIG emphasize they have not yet seen evidence of fundamentally new attack capabilities emerging from AI use.[42] While proofs of concept exist for LLM-powered malware that can modify its code and act autonomously, and threat actors are beginning to trial them in real-world attacks, they do not appear to yet have had much, if any, real-world impact.[43]

Admittedly, the overview in this section may not be comprehensive, as public reporting has often lagged nation-states' development and weaponization of sophisticated technological capabilities. In this case, however, the lag is probably smaller than with previous information technologies, because recent AI innovation has been driven primarily by the private sector.[44]

On the defensive side, leading cybersecurity companies report that LLM-powered tools are helping cybersecurity professionals with analysis and interpretation.[45] Microsoft reported that access to its Security Copilot tool significantly improved security professionals' accuracy, quality, and speed in key tasks.[46] Even so, as of March 2024, U.S. Department of Defense Senior Information Security Officer David McKeown was underwhelmed by AI-powered cyber defenses, saying, "I've been searching for use cases where AI is being used to do cybersecurity things and, so far, I'm not seeing too many."[47] Similarly, a July 2024 pilot study by the Cybersecurity and Infrastructure Security Agency (CISA) on AI-enabled vulnerability detection found that "[t]he best use of AI for vulnerability detection currently lies in supplementing and enhancing as opposed to replacing, existing tools; [i]n some cases, the amount of time needed for analysts to learn how to use the new capabilities is substantial and the incremental improvement gained may be negligible; and [i]n some cases, AI tools can be unpredictable in ways that are difficult to troubleshoot."[48]

Overall, while frontier AI systems' cyber capabilities are rapidly advancing, their practical impacts remain largely incremental for now. Attackers continue to derive results from less sophisticated attacks, and overburdened defenders are unlikely to adopt sophisticated AI tools unless these tools offer clear, ready-to-use benefits. Nevertheless, both adversaries and security teams are watching developments closely: Curious attackers are experimenting with the applications of new AI models as they are released, and defenders are eager for solutions that are both reliable and immediately deployable.

AI IN CYBERSPACE: OFFENSE AND DEFENSE

A COMPREHENSIVE ASSESSMENT of the national security implications of AI-powered cyber capabilities requires carefully balancing their offensive risks with their defensive potential. Overly restrictive regulation could inadvertently hinder defensive applications offering security benefits that exceed the risks posed. On the other hand, frontier AI has some qualitative distinctions that could challenge baseline assumptions that their diffusion as dual-use cyber tools should be expected to help defenders on balance.

Cyberspace offers attackers significant advantages: the complexity and interconnectedness of computer networks create large attack surfaces. Empirically, most modern software is "dense" with vulnerabilities: It is currently impractical to find and fix them all (and doing so may be computationally intractable for sufficiently complex software). Rather than aiming for 100 percent vulnerability-free systems, defenders can only hope that by discovering and addressing the most critical vulnerabilities, whatever they miss is sufficiently *difficult* for attackers to find and exploit. This asymmetry benefits attackers—finding and exploiting a single vulnerability is typically easier than patching them all.[49]

Moreover, diplomatic efforts to promote restraint face significant headwinds. Most cyber operations are inherently covert, complicating efforts to verify any country's commitment to restraint. Even when intrusions are detected, discerning intent is difficult, as cyber intrusions can have defensive value (for example, by providing insights into other countries' capabilities and intentions). Determining who is responsible for intrusions can also be challenging, and even where a given actor is found responsible, proving whether they had government backing and approval adds another layer of challenge.[50] Consistent with these challenges, U.S. cyber posture has evolved toward a more aggressive doctrine of persistent engagement. Formalized in Cyber Command's 2018 Command Vision, this strategy emphasizes that U.S. cyber operators should maintain a persistent presence inside adversary networks and proactively disrupt cyber threats before they can be launched against American targets.[51]

Defenders can only hope that by discovering and addressing the most critical vulnerabilities, whatever they miss is sufficiently difficult for attackers to find and exploit.

The United States stands to benefit from widespread development and diffusion of AI-enabled cyber defense capabilities. As a wealthy, globally connected, and technologically advanced country, the United States is an attractive target for cybercriminals and

state-backed cyber actors. Moreover, cyberattacks on allies and partners also impact U.S. economic and security interests—underscoring the benefits of improved cyber defenses that extend beyond U.S. borders. While improving AI-enabled cyber defense worldwide may challenge U.S. offensive cyber operations, U.S. capabilities have long been world-leading. As long as U.S. cyber warriors continue their track record of innovation and adaptation, this offensive edge will likely endure.[52]

AI's Potential for Defenders

Cyber defenders are well placed to benefit from advancing AI capabilities, even if those capabilities are also available to attackers. AI's ability to operate at machine speed and analyze data at scale can help defenders counter two of their most persistent disadvantages: the speed of automated attacks and the overwhelming scope of modern networks. AI's speed and scalability can both help defenders offset the otherwise significant structural advantages of attackers in the cyber domain.

Speed. In cybersecurity, attackers have traditionally maintained a significant temporal advantage over defenders. When intrusions occur, payloads can often propagate through networks at machine speed while human defenders are still processing initial alerts. By the time organizations identify and remediate a breach, attackers have frequently already achieved their goals. AI offers a promising path to help neutralize this temporal advantage. AI-powered security systems can support real-time detection and response capabilities that match the speed of automated attacks, dramatically reducing the critical dwell time between initial compromise and detection.

Scale. One of the most daunting problems for defenders is the sheer volume and complexity of their environments: Networks span continents and data flows at all hours. Traditional manual monitoring and rule-based alerting struggle to keep pace with attacks that can happen anywhere, at any time. AI can help turn this scale problem on its head. Machine learning–based defensive systems can simultaneously analyze billions of data points

One of the most daunting problems for defenders is the sheer volume and complexity of their environments.

around the clock, identifying subtle patterns and anomalies that would elude even well-resourced security teams. Moreover, defenders have opportunities to pool threat intelligence and training data—an advantage often unavailable to attackers who must operate secretly. In addition to AI's ability to handle scale, the scalability of the tools themselves also works in defenders' favor: Widely available AI systems for tasks like vulnerability discovery can be adopted across the broader defender ecosystem, making it more feasible to identify and patch issues at scale—even across the otherwise fragmented landscape of defensive actors.

Challenges to AI's Defensive Potential

While the previous discussion suggests that AI-enabled cyber capabilities will tend to favor defenders on the margin, this dynamic could plausibly shift in the coming years. The conceptual and empirical case for AI's defensive benefits remains strong—AI's capacity to enhance defenders' speed and scale represents a meaningful counterweight to attackers' traditional structural advantages. However, continued technological advancement along current trajectories might create conditions where any net defensive benefit provided from frontier AI capabilities is eroded or reversed.

This section analyzes three challenges that may lead to AI disproportionately empowering attackers as frontier AI progresses. This report dedicates comparatively more space to exploring this possibility not necessarily because the evidence for it is stronger, but because its actualization would represent a significant and destabilizing shift in the offense-defense dynamic with significant implications for U.S. national security.

First, growing inference costs at the frontier of capabilities may benefit well-resourced attackers who can selectively target high-value assets while defenders struggle to protect their entire attack surface. Second, automating the full cyber kill chain could compress operations from human to machine

speed, dramatically enhancing the potential of cyberattacks to support military and geopolitical objectives. Third, persistent technical challenges in model safety and reliability create asymmetric advantages for attackers with higher risk appetites who can better tolerate both system failures and collateral effects from their operations.

These technical challenges are further amplified by two cross-cutting factors. First, they will intersect with broader sociotechnical challenges countries and organizations face in integrating AI more deeply into their cyber defenses. Second, commercial and geopolitical competitive pressures will push AI developers to rapidly develop and deploy AI systems, at the potential expense of identifying and mitigating offensive risks.

GROWING INFERENCE COSTS AT THE FRONTIER OF CAPABILITIES

The defense-favoring status quo described earlier faces a potential challenge: accessing the most capable AI performance increasingly requires substantial spending on running models. Unlike traditional software, which typically has negligible running costs, using advanced AI tools increasingly involves nontrivial spending on inference compute. This means the cost of deploying sophisticated AI security tools could scale directly with the size of the attack surface, undermining one of the key advantages automated systems typically offer defenders.

If running state-of-the-art AI tools requires substantial computational resources, defenders may find it prohibitively expensive to protect their entire attack surface, while attackers may still be afforded success through selectively targeting smaller portions.[53] If the most powerful cyber capabilities increasingly hinge on inference spending, the economic dynamics of AI-driven cyber operations could begin to resemble those of human operations—rather than the defense-friendly economics of automated tools with more negligible scaling costs. While one might hope that the vulnerability discovery capabilities of AI systems will hit a ceiling as they become more powerful, in practice, the effectively inexhaustible number of vulnerabilities in sufficiently complex software systems (under prevailing development paradigms, at least) suggests that more powerful tools will continue to yield returns into the foreseeable future.[54]

The extent of this dynamic will ultimately depend on how strongly future state-of-the-art AI tools benefit from increased inference spending, and will be most relevant for motivated, well-resourced attackers. But overall, it cannot be taken for granted that increasingly powerful AI systems will inherently favor defenders, especially absent fundamental transformation of the software development process. This possibility warrants serious consideration given current trends in model capabilities and costs. While prices to access models of a given capability level have been consistently in free fall—decreasing at rates of at least 10 times annually since GPT-3's release in 2020—leading developers nonetheless continue to launch their most advanced models at price points in the tens of dollars per million output tokens. (A token is approximately four characters or three-quarters of a word.)[55]

If the most powerful cyber capabilities increasingly hinge on inference spending, the economic dynamics of AI-driven cyber operations could begin to resemble those of human operations.

Moreover, current frontier models—especially reasoning models—are able to perform better by running longer, seeing consistent benefits even into hours of runtime.[56] For the RE-Bench benchmark, which evaluates realistic machine learning engineering tasks, leading frontier AI systems are able to outperform human experts within the constraints of a two-hour window.[57] While the absolute cost of achieving certain tasks may decrease, the ability to convert financial resources into real-world performance advantages is increasing under current trends.

AUTOMATING THE FULL CYBER KILL CHAIN

Perhaps the most destabilizing development to the cyber offense-defense balance would be the rise of fully automated cyberattacks—throughout the whole kill chain, from reconnaissance through impact. By accelerating cyberattacks from human to machine speed, this automation could dramatically enhance the ability of cyberattacks to support military and geopolitical objectives.

Generally, cyberattacks are tactically fast but operationally slow. While the execution of attacks happens at machine speed, the planning, reconnaissance, and development of exploits remain labor-intensive and time-consuming processes. The most sophisticated operations often require months of preparation and thousands of hours of labor. Consequently, some of the most useful AI systems for U.S. offensive cyber operations have been those with broad capabilities to support campaign planning and strategic decision-making, rather than powerful but narrow capabilities.[58]

These prolonged preparation times hinder states from using disruptive cyberattacks to achieve military and geopolitical impacts—such as crippling systems during a military operation or coercing adversaries through attacks on critical infrastructure. To achieve impacts at the desired time, attackers generally will have needed to undertake advanced efforts to identify and prioritize key targets, successfully penetrate them, and then sustain this access. But in practice, given the inherent unpredictability of future operational needs, these preparatory efforts will often prove unnecessary or fail to provide operational value at the appropriate time. This unpredictability limits the practicality of disruptive cyberattacks for achieving military and geopolitical objectives, helping explain why, for example, Russia's cyberattacks played only a modest role in its invasion of Ukraine.[59]

If AI systems significantly reduce the time required to move from identifying targets to successfully executing a cyberattack, disruptive and destructive cyber operations would become both more attractive and practical as complements to conventional military actions—with potential destabilizing effects to the extent this bolsters the viability of military first strikes. At sufficiently compressed operational tempos, especially if defenders have similar abilities to "hack back," cyber conflicts could increasingly outpace the viability of human efforts to de-escalate and deter. When facing an escalating, fully automated cyber offensive, a defending state that does not deploy automated defenses could suffer substantial and irreversible harm in the time it takes for humans to navigate diplomatic channels or develop strategies to manage escalation and de-escalation.[60]

The emergence of fully automated sophisticated cyber operations would usher in a period of accelerated and disruptive change. Although early stages of automation are yielding incremental productivity gains, the most profound impact is likely to come from automation of the last remaining, most challenging steps in executing complex cyber campaigns. In practice, this will not happen all at once: progress will vary across different kinds of attacks and targets, depending on the specific demands of the operation and the defenses in place.[61] But as these harder-to-automate elements begin to fall away—across a growing range of scenarios—sophisticated operations will increasingly be freed from human bottlenecks, able to unfold at machine speed and scale.

Beyond accelerating operational tempo, automating the full kill chain could transform cyber operations in another crucial way: AI systems with meaningful autonomous capabilities would make it far more viable to wage sustained campaigns in highly secured networks. Today, cyber attackers often maintain ongoing communication links to manually direct and adapt their operations—such as Russian operators remotely accessing Ukrainian power grid systems to trigger outages, or China's Volt Typhoon moving through U.S. critical infrastructure networks, issuing commands to harvest credentials and conduct reconnaissance.[62] This dependence on outbound communication creates opportunities for defenders to spot malicious activity and, in some cases, to wrest control of command-and-control infrastructure.[63] In the Russian compromise of SolarWinds' Orion software, for example, identifying and ultimately redirecting attacker communications proved central to revealing the scope of the breach and neutralizing further spread.[64] The most sophisticated operations have gone to great lengths to sidestep this vulnerability entirely: Stuxnet's ability to execute sabotage inside Iran's air-gapped nuclear facility with minimal external communication required years of intelligence collection on specific systems and custom engineering for pre-programmed attack sequences.[65] Sufficiently adaptable AI-enabled systems could make such autonomous operation far more achievable—reducing or even eliminating the need for ongoing communication between attacking systems and their human controllers. While the transformative potential of automating the full kill chain is difficult to overstate, so too is the level of progress in capabilities that will be required.

Although current AI tools can boost human productivity and excel at specific tasks, realization of an "autonomous hacker" able to execute the entire process of a cyberattack remains speculative. Current AI systems' agentic abilities remain nascent, especially in the complex tasks required for sophisticated cyber operations.[66] Compared with humans, current AI systems generally struggle with long-term strategizing and maintaining coherent, productive work for extended periods, often losing focus or repeating unproductive actions.[67] Nonetheless, progress on easier tasks has sometimes been rapid, with newly released models performing respectably on agentic benchmarks where previous models rarely succeeded.[68] Overall, METR has found that the proficiency of leading AI models at autonomously completing longer and more complex software tasks—including in cybersecurity—has been increasing consistently and exponentially. The length of tasks they can complete has been doubling approximately every seven months over the past six years (on average, benchmarked against human experts). If this trajectory persists, by 2030, AI systems will be able to independently execute many software projects that would take human experts weeks or months to complete.[69]

Recent developments suggest that significant progress in agentic cyber capabilities may be within reach through building upon foundation models with subsequent large-scale reinforcement learning. In reinforcement learning, models learn from the success of their actions, rather than merely from existing data. On its own, large-scale reinforcement learning has achieved impressive breakthroughs in gaming, most notably when Google's AlphaGo defeated the world's top Go players in the 2010s—widely considered a landmark breakthrough in AI capabilities. Reinforcement learning systems have also approached human-level capabilities in real-time strategy video games such as StarCraft and Dota, which share interesting parallels with cybersecurity—both require operating with limited information about opponents' actions.[70]

Reinforcement learning is best suited to domains such as mathematics, where accurate solutions are objectively verifiable, allowing for direct reinforcement of behaviors that lead to correct outputs. In many ways, cybersecurity is such a domain: Success

Recent developments suggest that significant progress in agentic cyber capabilities may be within reach through building upon foundation models with subsequent large-scale reinforcement learning.

in capture-the-flag challenges, for example, is objectively verifiable. However, cybersecurity poses significantly greater challenges for reinforcement learning than gaming environments. While games operate within clearly defined rules and limited action spaces, cyber operations must navigate vast networks with countless possible configurations and attack vectors. Furthermore, games provide immediate, clear feedback through scores or territory gained, making it easy to measure progress. In cyberspace, it is far more difficult to automatically determine whether specific steps in an intrusion advance toward the intended goal. As a result, pure reinforcement learning approaches to developing AI models with broad, adaptable cyber capabilities appear impractical for now. Recent research—in particular, OpenAI's o-series models—has achieved strong results by leveraging foundation models as a starting point for subsequent reinforcement learning training, rather than relying on reinforcement learning alone.[71] This strategy could help unlock autonomous cyber capabilities that have remained out of reach.

Today's leading AI models already possess key building blocks: basic reasoning abilities, knowledge of common programming languages and cybersecurity concepts, and natural language processing. While these capabilities alone are not enough for autonomous cyber operations, they provide a far more sophisticated foundation for reinforcement learning training than starting reinforcement learning training from scratch. Furthermore, these existing capabilities can help with automatic generation of diverse training challenges and evaluation of model performance. Their natural language processing abilities can help parse network responses and error messages: gauging whether attempted intrusions are making progress, whether operational security is being maintained, and when systems have gone off track.[72]

Indeed, cutting-edge foundation models have shown promising abilities to generalize toward broad computer use skills through further training. When Anthropic trained its Claude 3.5 Sonnet model for computer use, the organization's researchers were "surprised by how rapidly Claude generalized from computer-use training on just a few pieces of simple software."[73] Despite being trained on basic tools such as calculators and text editors, Claude 3.5 developed respectable capabilities to use unfamiliar software without specific training. Further progress in AI models' ability to intelligently use a range of tools and approaches, even without specific training on them, would be a boon to cyber attackers and defenders alike. In DARPA's Cyber Grand Challenge, for example, teams found that higher-level autonomy—the ability to intelligently select and deploy different approaches based on circumstances—was often more crucial for systems' success than merely possessing particular capabilities.[74]

Time will tell whether building on foundation models with large-scale reinforcement learning will prove sufficiently scalable and effective to unlock sophisticated autonomous cyber capabilities. But recent advances in foundation models have significantly accelerated the rate at which AI models can develop general autonomous capabilities through training—a marked improvement from just a few years ago.

PERSISTENT TECHNICAL RELIABILITY CHALLENGES

As AI models become more capable at cyber operations, defenders and responsible actors face several technical challenges in ensuring their safe and controlled deployment. Three issues stand out: the difficulty of *assessing* AI models' true capabilities, the challenge of *constraining* those capabilities once discovered, and the *unreliability* of otherwise capable systems.

Assessing frontier AI model capabilities remains challenging. Researchers and practitioners can often elicit significantly improved performance from models through techniques such as providing access to additional tools, structuring inputs differently, fine-tuning, or deploying multiple copies of a model.[75] Without undertaking such efforts, assessments are liable to underestimate what a model is truly capable of. For example, Google researchers found they were able to improve LLMs' performance on a vulnerability discovery challenge by a factor of 20 by equipping them with standard coding tools and deploying the model in a framework that facilitated iterative exploration and self-correction.[76]

Techniques to constrain models fall short. Even when concerning capabilities are identified in AI models, reliably mitigating or controlling them remains an unsolved technical challenge. Current approaches include fine-tuning models to avoid prohibited outputs—for example, to comply with applicable laws and avoid enabling serious harm—and implementing additional moderation models on top of them to intercept such outputs.

While some fine-tuning techniques have shown promise, no approach has yet shown itself to be robust, with even moderately sophisticated adversaries proving able to circumvent restrictions.[77] The most aggressive attempts to suppress specific capabilities often lead to collateral degradation of related functions, making it difficult to target harmful behaviors without unintentionally impairing beneficial ones.[78] In short, there is not yet the ability to implement robust and reliable safeguards.

Models are still unreliable (even capable ones). Frontier models are capable of elegantly solving novel, extremely complex coding challenges and outperforming PhDs in knowledge of their domain of expertise.[79] At the same time, frontier models regularly stumble over basic logic puzzles, hallucinate false facts, or deviate from their training and instructions when given seemingly nonsensical prompts and inputs.[80]

This unreliability is more tolerable for attackers, particularly those with a higher risk appetite. For defenders, adopting unreliable systems risks disrupting their business or other activities. Attackers, on the other hand, will often be able to tolerate failures from unreliable offensive systems to realize significant upside when they do work.[81]

Moreover, less responsible actors likely have greater tolerance for potential collateral impacts from unreliable offensive cyber capabilities. Deploying powerful but unreliable autonomous systems could lead to attacks that may be tactically effective but

For defenders, adopting unreliable systems risks disrupting their business or other activities. Attackers, on the other hand, will often be able to tolerate failures from unreliable offensive systems.

come with heightened risks of miscalculations or unintended downstream effects. Historical (non-AI-enabled) examples in the cyber domain illustrate this risk. Stuxnet, a cyberweapon designed to sabotage Iran's nuclear program, spread beyond its intended target despite attempts by its developers to limit its reach, causing modest disruption outside of Iran as affected organizations investigated and remediated infections.[82] The NotPetya malware, initially designed and deployed by Russia against Ukrainian systems, spread globally, causing indiscriminate economic losses amounting to billions of dollars.[83]

Adversarial attacks—deliberately crafted inputs to exploit AI and machine learning systems—further exacerbate these model reliability challenges.[84] In the context of cybersecurity, where adversaries constantly seek to compromise systems, these attacks are virtually guaranteed when AI systems are employed. The risks of adversarial attacks go beyond AI systems merely failing to identify malicious activity. Just as biological immune systems can be exploited by autoimmune diseases that essentially turn the system against its host, automated cyber defense systems can also become an attack vector. For instance, attackers could exploit AI systems designed to quarantine network segments or restrict user privileges upon detecting suspicious activity. Such systems, if manipulated, might deny legitimate access to defenders, effectively weaponizing the defenders' tools against them. The more authority automated defenses have, the more they themselves become potential vectors for attack if compromised or manipulated.[85]

Alignment—the reliable embedding of specific goals and values within AI systems—poses a more fundamental and as yet unresolved challenge.[86] The alignment challenge requires those deploying advanced models to look beyond risks from models' mistakes or exploitation, accounting also for the risks that models themselves pursue unintended objectives. This challenge may be particularly acute in the cyber domain, given that adversarial capabilities such as deception and misdirection (or at least, understanding of them) are often desirable. For example, researchers found that fine-tuning LLMs (such as OpenAI's GPT-4o) to generate insecure code without disclosing this to users caused the models to develop broader malicious tendencies (such as encouraging violence and self-harm or admiring Hitler and Stalin).[87]

Researchers have also found that leading LLMs will sometimes behave deceptively to achieve their goals (especially when put under pressure to do so)—such as knowingly generating false outputs, attempting to disable oversight mechanisms or covertly copy themselves, or faking compliance during evaluations when they perceive their values would otherwise be modified. These behaviors were not explicitly trained for: Rather, the models determined on their own that deception was a useful strategy to accomplish their goals.[88]

The challenge of alignment may require the people overseeing the deployment of frontier models to account not only for models' mistakes or risk of exploitation, but also the possibility that the models themselves are inadvertently trained with inherent goals conflicting with or at cross-purposes to the intentions of their human designers. As developers and customers pursue models capable of strategizing to achieve goals, particularly those trained in deceptive and adversarial behaviors, they must recognize that models' alignment with developers' intentions cannot be taken for granted.

All the preceding technical challenges will interact with other emerging technological dynamics. For example, one concerning risk is the potential rise of automated front-running of security patches.[89] Currently, when patches are released, attackers typically need days or weeks to reverse-engineer them and develop exploits—although the time-to-exploit period has trended downward in recent years.[90] Increasingly rapid and scalable exploitation of vulnerabilities will pressure defenders to move in tandem toward automated, real-time patching. But this is a risky proposition, given the potential for system-wide disruptions from flawed patches, as demonstrated by the July 2024 CrowdStrike incident.[91] The rapid

evolution of frontier AI is outpacing humans' ability to reliably evaluate, constrain, or align it. If unaddressed, this will increasingly leave defenders disadvantaged against attackers willing to accept higher risks.

SOCIOTECHNICAL CHALLENGES

Fully harnessing AI's defensive potential is not just a technical challenge: Organizations must also take care to not mismanage human-machine interactions through overreliance on automation, inadequate training, or insufficient safeguards. Failure to do so will risk undermining the defensive advantages of these systems even as the underlying technology improves.

Relying on automated systems in domains like cybersecurity poses fundamental challenges: automation bias, de-skilling, and eroded sensitivity to operations. Automation bias—the tendency to place excessive trust in automated outputs—has been a contributing factor to catastrophic failures in high-stakes environments, such as fratricides from missile defense systems. De-skilling occurs as operators lose crucial expertise through prolonged reliance on automation. Eroded sensitivity to operations arises as operators' real-time understanding of automated systems diminishes in the face of a reduced need for active engagement. When a highly automated system encounters widespread failure or novel adversarial tactics, human operators are often forced to intervene without the benefit of regular hands-on experience or a comprehensive understanding of the system's current state.[92] These risks are particularly acute in systems that are either highly centralized—a tempting default, given the economies of scale that characterize the development and deployment of AI—or tightly coupled (that is, where subcomponents are quickly and directly responsive to each other): In these systems, initial failures can propagate widely.[93] Building resilient systems therefore requires maintaining personnel capable of manual intervention, diagnostic reasoning, and flexible problem-solving.

These human-machine interaction challenges are compounded by broader organizational and systemic risks from AI integration. While AI tools offer powerful centralized and scalable capabilities, their widespread adoption introduces systemic risks—vulnerabilities in commonly used models can have outsized effects, and interconnected systems may interact in unexpected ways. A robust cyber defense strategy—whether at the level of organizations or nationally—must include diverse techniques for detecting and frustrating attacks, thoughtful partitioning of responsibilities between various tools and people, contingency planning for system failures, and sufficient investment in maintaining human expertise.

Countries and organizations frequently face periods of heightened vulnerability while transitioning to new capabilities as they adapt their systems, protocols, and behaviors.

The dynamics of strategic competition in cyberspace will further amplify these challenges. Jacquelyn Schneider's concept of the "capability-vulnerability paradox" highlights how enhanced capabilities often create new vulnerabilities—resources and technologies that increase operational effectiveness can themselves become critical weaknesses. Countries and organizations frequently face periods of heightened vulnerability while transitioning to new capabilities as they adapt their systems, protocols, and behaviors. These transition periods can create destabilizing first-strike incentives for large-scale attacks: Less capable actors might exploit temporary windows of vulnerability before stronger states fully integrate and realize the potential new capabilities, while stronger states face pressure to act preemptively during their periods of transitional weakness.[94]

COMPETITIVE PRESSURES

The conditions under which AI developers and deployers make decisions will also influence their ability to make informed, prudent choices around developing and disseminating AI-enabled cyber capabilities.

Domestic competitive pressures. AI developers are under intense pressure to rapidly develop and deploy advanced AI systems, with billions in revenue and trillions in market capitalization hanging in the balance. To attract the highest levels of investment, developers need to position their models

as state-of-the-art and demonstrate strong user demand.

Leading U.S. developers currently self-assess the risks their models pose. In their respective safety frameworks, both Anthropic and Google DeepMind acknowledge that competitive pressures could lead them to forgo certain risk mitigations they would otherwise consider justified, if competitors fail to implement similar mitigations for similarly capable models.[95] Recent model releases have highlighted the tension between the push to quickly ship products and to surface potential threats. Through most of 2024, when leading frontier AI developers released new, more powerful models, they typically accompanied these releases with technical papers; these papers included the results of evaluations of the models' capabilities and risks across various domains, including offensive cyber operations.[96] However, there has been an apparent trend toward compressing these evaluations and/or delaying or omitting their public release:

- After OpenAI's September 2024 release of its then-state-of-the-art o1-preview model, one of the external safety reviewers, METR, noted challenges in evaluating the full capabilities of the model. METR was only given six days, in which time they were able to elicit substantial performance gains from the model by iterating on its agent scaffold (i.e., the framework the model operated within and the tools it had access to). They concluded there was still "substantial room for improvement" at the end of their evaluation period.[97]
- Most recently with its flagship Gemini 2.5 Pro, Google DeepMind has made its most powerful models available to the public for months before releasing detailed evaluations of their cyber capabilities.[98]
- In February 2024, OpenAI released its then-state-of-the-art o3 model-powered Deep Research tool three weeks before releasing a risk-focused technical paper.[99]
- For OpenAI's o1 and o3 models above, evaluations were conducted on nonfinal versions of models, with one OpenAI researcher explaining that, "making a 100 page report on preparedness is really time consuming work that has to be done in parallel with post training improvements."[100]
- In July 2025, xAI released its then-state-of-the-art Grok 4 model more than a month before releasing associated safety documentation.[101]

Given strong early indications that a model poses tolerable real-world threats, these decisions are not necessarily inappropriate—especially as closed-weight models can be recalled if unforeseen patterns of misuse emerge. Moreover, these developers have established arrangements with the U.S. Center for AI Standards and Innovation (CAISI, formerly the AI Safety Institute) and/or UK AI Security Institute (formerly the UK AI Safety Institute) to provide them with pre-release model access.[102] Nevertheless, it does appear that frontier AI developers are increasingly facing tradeoffs between their initial risk-focused model release practices and competitive realities.

International competitive pressures. Tensions between managing competitiveness and risk also exist at the international level. In particular, the accelerating race to deploy frontier AI models as a national security imperative will create strong obstacles to meaningful collaboration between the United States and China to mitigate related risks.

Efforts to understand and defend against AI-enabled cyber threats inherently overlap efforts to weaponize these capabilities. Given the degree of active cyber competition between the two countries, both will likely hesitate to be transparent about their domestic capabilities, limitations, safeguards, and threat modeling.

At the same time, there are modest areas of cooperation and attention to AI risks. China signed the Bletchley Declaration at the 2023 AI Safety Summit, which acknowledges cyber risks. China's National Technical Committee 260 on Cybersecurity—a key technical standardization body within China's national standards system—also released an AI Safety Governance Framework identifying cybersecurity as a key concern.[103]

MODEL RELEASE AND SECURITY CONSIDERATIONS

AS DEVELOPERS AND POLICYMAKERS NAVIGATE the above challenges, they will face decisions not only about the development of tools, but also on influencing deployment and access to these systems. How AI tools are released—and whether they can be adequately safeguarded if not widely released—will play a key role in determining their overall impact on the cybersecurity landscape.

The diffusion of benefits and risks from AI systems ultimately depends on how they are released. While broad distribution of AI-enabled cyber capabilities will often serve U.S. interests, immediate and unlimited release may not always be the best approach. In practice, this will depend on weighing the advantages of broad access for defenders against the risks from adversaries gaining access—and the extent that alternative release strategies can limit the latter without sacrificing too much of the former.[104]

The picture is further complicated when considering the downstream effects of releasing dual-use cyber tools. When a tool is made public, it doesn't necessarily follow that subsequent know-how and tools built on it will also become public. This is an important consideration for machine learning models: Unlocking their full potential often requires discovering fruitful strategies for interacting with them, giving them access to tools, or fine-tuning them on relevant data.[105] As AI models become more powerful, it will be important to account for the full range of possible release strategies. Phased distribution, for example, could allow particularly powerful tools to first be used to identify vulnerabilities in particularly critical codebases, giving time for developers to address newly discovered vulnerabilities before the tools are made widely available.

This raises an important empirical question: Do more powerful AI systems consistently find all vulnerabilities identified by weaker ones? Or is there enough variability that weaker systems can sometimes discover vulnerabilities that stronger systems cannot (as is the case with human security researchers)? The answer has practical implications for U.S. policy.

When a tool is made public, it doesn't necessarily follow that subsequent know-how and tools built on it will also become public.

If more powerful systems reliably catch everything that weaker ones can, the United States faces fewer risks from sharing its AI cybersecurity tools—as long as it maintains access to world-leading systems, adversaries using similar or weaker systems will have limited ability to find vulnerabilities the United States cannot. But if weaker systems have abilities to find unique vulnerabilities, the risks of widespread

diffusion become more acute. Even as the United States maintains overall technological superiority, adversaries could integrate and build upon U.S. tools and insights within the quirks of their own ecosystem, potentially discovering vulnerabilities that U.S. systems alone systematically overlook.

As AI models become more strategically important, they will also become more attractive targets for adversaries to exfiltrate—creating a self-reinforcing cycle where AI advancements make them both more valuable and vulnerable. Leading foundation models reflect immense investments in research, engineering, and computation, distilling vast knowledge into a few terabytes of ready-to-deploy intelligence. These systems, replicated across thousands of instances, are increasingly seen as strategic assets—coveted by adversaries who recognize AI as an enabler of global influence. Anthropic, for instance, acknowledges that its systems are not yet hardened against the concerted efforts of advanced state actors. (This admission likely reflects transparency on Anthropic's part, as opposed to the organization having worse information security practices than its competitors.)[106]

As the cyber capabilities of AI models grow more sophisticated, they may soon play a decisive role in government or private cyber operations. Indeed, the offensive cyber capabilities of AI models could become increasingly central to their strategic value. This trajectory could pose a challenge: The same advances that make AI models valuable for cyber offense will also make them prime targets for adversaries seeking to steal, compromise, or weaponize them. This dynamic creates a self-reinforcing cycle: As AI models advance, their protection becomes more vital yet more challenging.[107]

Compounding this challenge is the emerging possibility of models themselves actively participating in their attempted exfiltration. In certain scenarios, leading LLMs will sometimes demonstrate a willingness to covertly participate in their own exfiltration. Specifically, in simulated scenarios where models discovered they were likely to be modified or replaced with models that did not share their current goals or values, most will, given the opportunity, sometimes elect to try covertly copying their model weights to an external server.[108] These results signal a need for a cautious approach to the reliability of traditional methods for securing model weights: Developers should consider the possibility that sufficiently capable models could subvert security protocols, effectively becoming insider threats in their own right.[109]

CONCLUSION AND RECOMMENDATIONS

WHILE AI HAS HISTORICALLY HELPED blunt cyber attackers' structural advantages, emerging developments could tilt it toward bolstering attackers, with serious implications for national security. First, as inference costs at the frontier of AI capabilities grow, defenders may find it unaffordable to apply state-of-the-art tools across their whole attack surfaces, even as attackers can opportunistically target narrower footholds. Second, automating the full cyber kill chain could accelerate operations from months of human effort to machine-speed execution, increasing the military and geopolitical utility of cyberattacks while reducing opportunities for human de-escalation. Third, persistent technical reliability challenges will increasingly advantage attackers who can tolerate failure and collateral damage over defenders who cannot. All these challenges will unfold against a backdrop of intense domestic and international competition around AI.

But the net offensive or defensive advantage of AI-powered cyber tools is not predetermined. It will vary across tools and their capabilities, depend on decisions around development and deployment, and can be systematically assessed. The U.S. government has a crucial role to play—it must prioritize a deeper understanding of evolving AI-enabled cyber threats and proactively shape the technological landscape in favor of U.S. security and stability.

The report offers the following recommendations for policymakers to strengthen cyber defenses, invest in research that differentially empowers defenders, improve evaluation frameworks, enhance regulatory clarity and effectiveness, and promote security at frontier AI developers.

Double down on policies to shore up cybersecurity.

Although mitigating threats from AI-enabled cyber operations will require governing relevant AI systems, traditional cybersecurity practices remain as important as ever. Indeed, as AI accelerates potential cyber threats, the importance of foundational cybersecurity efforts will only increase. This requires the U.S. government to raise the bar for cybersecurity through security-focused procurement and stronger accountability for failures. Congress should also ensure that efforts to address AI-related risks do not come at the expense of broader cybersecurity priorities.

Organizations' cybersecurity often relies more on attackers' finite resources than on robust defenses—an increasingly untenable strategy in the face of escalating and scalable AI-enabled threats. While individual entities are responsible for their security, they cannot be expected to develop adequate defenses independently. To achieve greater security at scale,

technology providers must lay the foundations of cyber defense by adopting secure by design and secure by default practices: building and layering security features and safeguards from inception rather than as afterthoughts, and ensuring they are active by default.[110]

Secure by design and secure by default practices are not yet widespread among developers, as evidenced by the persistent prevalence of basic security flaws that should be prevented through proper design principles. In 2007, the nonprofit MITRE Corporation identified 13 "unforgivable" vulnerabilities: common, well-documented security flaws that were easily detectable and indicative of inadequate security awareness or testing practices. Alarmingly, most of these vulnerabilities persist today in MITRE's list of "stubborn weaknesses"—software vulnerabilities that have consistently featured among the top 25 most dangerous (based on frequency and severity)—suggesting a vast landscape of insecure software that could be exploited at scale with even modest advances in autonomous cyber agents.[111] Fortunately, AI is certainly not necessary to address these vulnerabilities—but it will likely have an important role to play, given the scale of the challenge.

As this potential precipice nears, the U.S. government will play a central role in determining societal tolerance for this degree of insecurity—as it sets standards, makes procurement decisions, and determines policy settings for accountability in the aftermath of security breaches. As long as the current pace of AI progress continues, Washington should become less tolerant of cyber insecurity.

More broadly, Congress should not let its AI-related efforts sideline its broader responsibilities on cybersecurity. Progress has stalled, for example, on adopting recommendations of the Cyberspace Solarium Commission, including around aspects of federal incident response capabilities, scaling up of information sharing, and developing certifications for cloud security and cyber insurance.[112] Given the potential for AI to significantly reshape the cybersecurity landscape, Congress should ensure that CISA maintains the strategy, capacity, and competencies needed for effective coordination.[113] It should also redouble efforts around regulatory harmonization to free up defender time to integrate and leverage new capabilities in the face of attackers doing the same.[114]

Invest in AI research and development to differentially promote cyber defense.

The United States is well-placed to benefit from improving AI-enabled cyber defenses. Even though U.S. companies such as Microsoft and Google have invested in the development of frontier AI-enabled tools to support defenders in cyberspace, there is a compelling case for the U.S. government to provide complementary support.

The U.S. government plays a critical role in supporting foundational, pre-commercialization research. Federal research agencies such as the National Science Foundation (NSF), DARPA, and Intelligence Advanced Research Projects Activity (IARPA) should fund high-risk, high-reward projects to unlock fundamental advances in empowering cyber defenders through powerful AI models. Areas of focus could include the following:

- ***Robustness:*** Despite their remarkable abilities, current cutting-edge AI systems are surprisingly brittle in the face of unexpected inputs or deliberate attempts to mislead them—a particular liability for cyber defenders (see page 13). This focus is consistent with the recommendations of the 2025 AI Action Plan.[115]

- ***Privacy-preserving multiparty computing:*** Ideally, cyber defenders would combine relevant data to get a fuller picture of emerging threats, but security and privacy concerns often prevent this kind of collaboration. Secure multiparty computation techniques offer ways to analyze shared data without exposing it, but existing methods are largely too slow and complex to prove practical. Advancing their efficiency and scalability would better enable defenders to leverage their collective scale.[116]

- ***Formal verification at scale:*** Current formal verification methods require specialized expertise and intensive manual effort, limiting their adoption beyond high-stakes environments such as avionics or cryptographic libraries—leaving most software vulnerable to security flaws. AI-assisted tools could help automatically verify security properties such as memory safety and access controls across far more systems, dramatically reducing the attack

surface that defenders must protect. While the challenge of precisely defining security specifications is hard to overstate, advances in automated theorem proving and more intuitive specification languages could enable mathematical security guarantees to play a larger role in the broader software ecosystem.[117]

The U.S. government also plays a vital role in supporting standards development through the National Institute of Standards and Technology (NIST). In the face of a potential surge of AI agents in cyberspace, NIST and CAISI should develop and promote cybersecurity standards:

- standards to support synthetic content provenance and authentication
- standards for promoting security while facilitating AI agents' interaction with IT systems
- standards for identity and authentication to verify human users and distinguish specific autonomous agents[118]

More broadly, the U.S. government has a role to play in investing in the security of the cyber commons that may not otherwise receive security attention commensurate with its importance. For example, society relies heavily on open-source code, often maintained by small teams of volunteers or with limited resources. Vulnerabilities in this widely used code can have severe, potentially devastating, consequences. For example, the Heartbleed bug in the crucial OpenSSL encryption library—responsible for securing much of the internet's traffic—exposed sensitive data; similarly, the widespread Log4j vulnerability (a logging utility embedded in countless applications) required a massive, global patching effort to prevent widespread system compromise. Federal research and cybersecurity agencies such as the NSF, DARPA, and IARPA should continue seeking promising AI-enabled opportunities to improve security. Examples include helping automate the porting of code to memory-safe languages (as in DARPA's Translating All C to Rust [TRACTOR] efforts), better aligning LLM code generation to secure coding practices, and improving the efficiency with which known vulnerabilities can be addressed.[119]

Strengthen evaluation of AI cyber capabilities and risks.

To ensure U.S. national security amid rapidly advancing AI capabilities, the U.S. government must maintain independent awareness of emerging cyber capabilities associated with frontier models. Foreign and open-weight models now lag the U.S.-dominated AI frontier by mere months. Failure to monitor progress in the foreign and open-weight ecosystems risks blindness to novel, AI-enabled cyber capabilities that could result in actual attacks on U.S. or allied networks.

Unlike past technologies such as nuclear weapons or space systems—developed primarily through government programs—the private sector is driving nearly all development of frontier AI systems. To address the associated information asymmetry, the U.S. government needs both robust evaluation capabilities and authority.

FORMALLY AUTHORIZE CAISI AND PROVIDE IT WITH STABLE, LONG-TERM FUNDING.

Even if the U.S. government delegates technical assessments to AI labs and third-party evaluators, it must form its own view about what constitutes sufficient and trustworthy evaluation of AI capabilities and their national security implications. Assessing the full extent of frontier models' capabilities and the robustness of their safeguards is not straightforward: One organization may succeed at eliciting capabilities from a model while another fails—and AI developers and deployers face conflicts of interest, given the immense and growing competitive pressures to deploy.[120] Moreover, the U.S. government possesses unique, nonpublic insights into cyber offense and defense dynamics, including adversary capabilities and tactics, that are critical for effectively identifying and mitigating emerging risks.

Recognizing this, the 2024 National Security Memorandum on AI designated CAISI as the primary federal point of contact for private AI developers to share voluntary pre- and post-deployment safety testing of frontier AI models, including comprehensive assessments of cybersecurity risks. In this capacity, CAISI develops and implements cyber testing protocols that evaluate AI models' abilities to detect, generate, and mitigate offensive cyber threats.

Additionally, CAISI is responsible for establishing and maintaining benchmarking standards to objectively measure AI models' cybersecurity capabilities and vulnerabilities. CAISI also coordinates with the National Security Agency's (NSA's) AI Security Center on classified testing and assessment of AI models' offensive cyber risks.[121] This role for CAISI is also acknowledged in the Trump administration's 2025 AI Action Plan.[122]

Congress should formally authorize CAISI and provide it with stable, long-term funding to support a robust research and development pipeline, attract top talent, and cement its credibility and position as a long-term institution.[123]

CODIFY THE U.S. GOVERNMENT'S AUTHORITY TO EVALUATE FRONTIER MODELS.

The U.S. government needs not only the *capacity* to evaluate frontier models for cyber capabilities; it also needs the *authority*. While leading frontier model developers' voluntary provision of access to CAISI and other government departments is commendable, relying on this voluntary arrangement is insufficient to protect U.S. national security. Congress should codify the right of a designated U.S. government body to undertake or compel evaluations of frontier models.

In September 2024, the Bureau of Industry and Security (BIS) proposed a rule requiring developers to inform the U.S. government about their efforts to train "dual-use foundation models" (currently defined as models trained using 10^{26} numerical operations). This information would include red-teaming results relating to "the discovery of software vulnerabilities and development of associated exploits; the use of software or tools to influence real or virtual events; the possibility for self-replication or propagation; and associated measures to meet safety objectives."[124] However, whether BIS will ultimately implement this rule is uncertain, as it was precipitated by President Joe Biden's now-revoked Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

The BIS rule's use of the Defense Production Act powers outside of the traditional defense industrial base drew some criticism.[125] But even so, the actual requirement was relatively limited: Merely requiring developers to share red-teaming results still leaves the extent and quality of that red-teaming to their discretion. To address these issues, Congress should codify the right of a designated U.S. government body (such as the Department of Commerce, given it already houses CAISI) to undertake or compel evaluations of frontier models that meet certain criteria, with appropriate safeguards around intellectual property, regulatory burden, and due process. These evaluations should also include a defined focus on large-scale national security risks.

While the use of a training compute threshold to target oversight has attracted criticism, this practice should not be abandoned. Increased compute remains a key driver of AI capabilities, and while capabilities from less compute-intensive models continue to improve, transformative cyber capabilities are likely to first emerge downstream of large training runs (see pages 5–7). Moreover, training compute is a straightforward metric that developers can anticipate. Although training compute is only one factor contributing to advancing AI capabilities, it remains a useful initial filter to determine which models warrant further evaluation, thus reducing the regulatory burden for most AI developers. Appropriate thresholds will change over time and should be updated accordingly.[126]

ENSURE THAT EVALUATIONS OF AI CYBER CAPABILITIES AND RISKS COVER THE RIGHT THINGS.

As the AI industry, wider research and evaluations ecosystem, and governments mature their efforts to evaluate and model AI-enabled cyber capabilities and risks, they should promote the following practices:[127]

- Contextualize evaluations in real-world threat models, along with existing frameworks for the cyber domain, such as MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge—see the appendix), so that cybersecurity practitioners can interpret their real-world implications.[128]
- Contextualize evaluations by comparing AI-enabled capabilities to baselines from humans, existing tools, and foreign AI models, including around factors such as cost and time taken, to determine the *marginal* impacts of AI systems.[129]
- Develop evaluations that will prove challenging for even extremely capable models, to make evaluations future-proof.[130]

- Promote feedback loops between evaluations, predicted effects, and real-world threat intelligence.
- Account for the net effect of capabilities that can support both cyber offense and defense.
- Mature capabilities to conduct secure, privacy-preserving evaluations (using technologies such as confidential computing), allowing robust assessments while safeguarding the privacy and security of developers' and evaluators' intellectual property.[131]

Evaluations should focus directly on cyber capabilities, but also on related factors that could impact the cyber offense-defense balance. These include:

- progress toward reliable, autonomous AI agents[132]
- the cost-effectiveness of deploying the most advanced capabilities at scale
- the variance in cyber capabilities across similarly cyber-capable models (for example, whether more powerful AI systems can consistently find the same vulnerabilities as less powerful systems)
- capabilities and propensities of the most capable models to subvert authorized human control
- how the effort required to find increasingly complex vulnerabilities scales toward the highest levels of difficulty

Additionally, researchers should assess the sustainability of key drivers of progress—such as algorithmic improvements, data, and compute—to determine how appropriate it is to extrapolate current trends in AI cyber capabilities into the future.

Sufficiently resource BIS to enforce AI-related export controls.

Training and inference compute remain central to advancing frontier AI capabilities, including in the cyber domain. To maintain America's AI edge, the U.S. government has imposed sweeping export controls targeting China's access to cutting-edge AI chips, as well as the associated upstream supply chain for manufacturing.

The rationale for these controls is sound. Like any country, China operates within budgetary constraints. Increasing its cost of compute disincentivizes the compute-intensive undertakings required for frontier AI models that may pose threats to U.S. security. This dynamic will only intensify as training costs for cutting-edge AI models continue to rise.

Despite claims otherwise, neither the impressive performance of DeepSeek's R1 model nor the growing role of inference compute undermines the case for the controls. Even as a shift to reasoning models and other efficiency improvements enable increasingly capable models at lower levels of training compute, compute scaling remains an important driver of state-of-the-art capabilities, including in the cyber domain (see pages 5–7). Moreover, growing spending on inference compute only heightens the impact on adversaries of increasing their cost of compute, including for chips that are otherwise less suited to training models.[133]

Export controls are only as effective as their enforcement.

But export controls are only as effective as their enforcement: Smuggling of cutting-edge chips is widespread, and Huawei and Semiconductor Manufacturing International Corporation (SMIC) have in some cases still been able to get their hands on advanced equipment integral to helping them advance their chip production capabilities (although January 2025 updates to export controls go some way toward mitigating this).[134] BIS's funding has been stagnant as its mission grows; single shipments of high-end chips exceed its enforcement budget. Congress should address BIS's chronic underresourcing, enabling it to move from reactive to proactive and modernized enforcement, consistent with the 2025 AI Action Plan.[135]

Clarify federal regulation around reasonable care and liability for cyber harms from frontier AI.

Congress should create a federal regulatory framework that clarifies reasonable care and liability for

cyber harms caused by frontier AI models. Such a framework would likely feature a combination of codified principles, designation of a federal regulatory authority, and some targeted preemption of relevant state laws. This framework is important both to prevent a patchwork of state laws that could stifle the beneficial uses of AI in cybersecurity and to ensure that negligence is appropriately deterred as AI-enabled cyber capabilities pose a growing threat to U.S. national security.

To be clear, state-level AI policies have a role to play. The House Task Force Report on AI highlights how state policies can help reflect local needs and allow for policy experimentation.[136] However, large-scale cyberattacks typically cross state (and even national) borders, necessitating a broader approach. Furthermore, the immense cost of training frontier models makes it impractical for developers to tailor training to individual state laws.[137] State-level policies impacting early-stage training will inevitably have national consequences. Federal coordination is therefore essential.

AI developers should not view federal clarification and codification of liability as necessarily imposing a regulatory burden. These federal actions can reduce uncertainty, account for important nuances, and create safe harbors (such as for cheaper or open-weight models). Given the rapid pace of AI progress, failure to build a federal regulatory capacity—with relevant expertise, powers, and well-thought-out decision-making processes—could just as easily result in overreactions—whether from the executive or legislature, or from the judiciary as it develops novel precedents.

Current cybercrime regulations generally do not hold developers and providers of offensive cyber tools liable unless they intentionally promote malicious use. This protects innovation and access to tools with defensive or non-cybersecurity applications. For *unintentional* enablement of cybercrime, however, negligence tort law is more relevant. Under negligence tort law, a claimant would need to prove that an AI developer or deployer failed its duty of care—the legal obligation to act as a reasonably prudent person would, considering factors such as foreseeable harm, the burden of precautions versus the risk, and broader public policy.[138] But in the absence of precedent relevant to the unique aspects of frontier AI, there is in practice uncertainty for both developers and potential plaintiffs.

Because frontier models are general-purpose software, the bar for liability should be high. Chilling their development would jeopardize a wide array of other benefits. Even so, certain safeguards are more viable for general-purpose frontier AI than for traditional software. LLMs can be trained and instructed around the type of activities they should and should not assist with (although the imperfect performance of these safeguards would make it premature to enshrine them in policy for now). Furthermore, frontier model providers often offer access solely through online platforms under their control. This allows them to use automated models to flag suspicious queries and to implement know-your-customer practices.

To mitigate the risks of overreach, this federal regulatory effort should adhere to the following principles:

- Apply only to frontier models.
- Have higher bars for regulating open-weight models.[139]
- Use scheduled reviews or sunsets to ensure requirements remain appropriate given progress in technology and the understanding of relevant risks.
- Require regulators to account for both the costs and benefits of restrictions.
- Require regulators to account for the marginal risk that models pose over already available tools and information—including foreign models.[140]
- Focus solely on large-scale harms to otherwise uninvolved parties, rather than model users and their customers (who have more remedies and opportunities to opt out).
- Aim to proactively signal the levels of cyber capabilities that will trigger regulatory oversight.[141]

Promote information security at frontier AI developers.

The importance of supporting the secure development and deployment of cutting-edge AI models will only grow with their capabilities. Frontier AI developers

and deployers should bolster their security commensurate with the growing commercial and national security value of their intellectual property, especially to sophisticated U.S. adversaries. To successfully defend their models, frontier AI developers will need support from the U.S. government.[142]

The intelligence community should closely monitor advancements in frontier AI models, particularly regarding their potential for weaponization by U.S. adversaries. Given the pace of AI progress, nascent information security practices within organizations, and the lead time required for security improvements, the U.S. government should evaluate whether and when it is necessary to intervene, such as by offering industry cybersecurity assistance through agencies such as CISA and the NSA.[143] Additionally, given that most frontier AI developers and their data center partners now contract with the U.S. government, agencies should prioritize information security as they procure frontier AI capabilities.[144] The intelligence community should also prioritize intelligence collection on potential threats to the security of leading models and developers.

Promote global norms around liability for harms from automated cyber operations.

As AI agents become increasingly capable of broad and sophisticated actions in cyberspace, there is a risk that developers and deployers may attempt to deflect responsibility for related harms. As bodies such as the U.S. Department of State's Bureau of Cyberspace and Digital Policy engage internationally around cyber norms, they should seek opportunities to promote the development and adoption of international norms (or law, eventually) to clearly define accountability for the actions of autonomous agents in cyberspace.

APPENDIX:

Stages of Cyber Operations and Example Applications of AI

Cybersecurity professionals generally think about both offensive and defensive operations in terms of multiple stages, with each offering opportunities to leverage AI capabilities. Recognizing these stages supports nuanced analysis:

- → It enables more precise analysis of AI's impact—advances may dramatically accelerate certain phases while leaving others relatively unchanged.
- → It highlights why cyber operations are generally "tactically fast but operationally slow"—individual engagements can execute at machine speed, but sophisticated operations typically require months of human planning and coordination.
- → It exposes the limits of isolated capabilities—transformative automation will demand systems that can coherently orchestrate multiple stages over time while dynamically adapting to discoveries and setbacks.

MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge) is one of the most widely adopted frameworks in cybersecurity, developed by the nonprofit MITRE Corporation to document and classify common adversary tactics and techniques based on real-world observations. It is widely used as a foundation for developing threat models and defensive methodologies in the cybersecurity community.

MITRE ATT&CK Framework and Example Applications of AI[145]

Stage	Description	Example Application of AI
Reconnaissance	The adversary tries to gather information it can use to plan future operations.	Scanning for vulnerable entry points online
Resource Development	The adversary tries to establish resources it can use to support operations.	Discovering and developing exploits for vulnerabilities
Initial Access	The adversary tries to get into the defender's network.	Developing and executing spear phishing campaigns
Execution	The adversary tries to run malicious code.	Selecting and using exploits for identified vulnerabilities
Persistence	The adversary tries to maintain its foothold.	Identifying opportunities to add authorized public keys
Privilege Escalation	The adversary tries to gain higher-level permissions.	Identifying opportunities to trigger privilege escalation
Defense Evasion	The adversary tries to avoid being detected.	Emulating existing network patterns
Credential Access	The adversary tries to steal account names and passwords.	Detecting passwords stored in plaintext
Discovery	The adversary tries to figure out the defender's environment.	Identifying network locations likely to be of interest to the attacker
Lateral Movement	The adversary tries to move through the defender's environment.	Hijacking remote desktop sessions

continued

Stage	Description	Example Application of AI
Collection	The adversary tries to gather data of interest to its goal.	Identifying sensitive information in system calls
Command and Control	The adversary tries to communicate with compromised systems to control them.	Identifying opportunities to include communications in otherwise legitimate network traffic
Exfiltration	The adversary tries to steal data.	Disguising exfiltrated data as legitimate network traffic
Impact	The adversary tries to manipulate, interrupt, or destroy the defender's systems and data.	Identifying high-impact opportunities to disrupt victim operations

Cyber defense also occurs across distinct functions.[146]

Function	Description	Example Application of AI
Prevention	Finding and patching weaknesses	Fuzzing and penetration testing
Detection	Discovering network activities of potential concern	Alert prioritization
Response and Recovery	Thwarting and addressing attacks	Automated isolation of potentially compromised machines
Active Defense	Aiming to proactively "engage or study external actors" through "a spectrum of activity that includes annoyance, attribution, or outright counterattack"	Dynamic creation of honeypots

1. *Navigating the AI Frontier: A Primer on the Evolution and Impact of AI Agents* (World Economic Forum, December 2024), https://www.weforum.org/publications/navigating-the-ai-frontier-a-primer-on-the-evolution-and-impact-of-ai-agents/; Helen Toner, et al., *Through the Chat Window and into the Real World: Preparing for AI Agents* (Center for Security and Emerging Technology, October 2024), https://cset.georgetown.edu/publication/through-the-chat-window-and-into-the-real-world-preparing-for-ai-agents/; and "Building Effective Agents," Anthropic, December 19, 2024, https://www.anthropic.com/engineering/building-effective-agents.

2. "Cyber Kill Chain," Lockheed Martin, https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html.

3. Bill Drexel and Caleb Withers, *Catalyzing Crisis: A Primer on Artificial Intelligence, Catastrophes, and National Security* (Center for a New American Security, June 2024), n. 89, https://www.cnas.org/publications/reports/catalyzing-crisis.

4. Yoshua Bengio, et al., *International AI Safety Report* (UK Department for Science, Innovation and Technology, January 2025), 221, https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf.

5. Charles L. Glaser and Chaim Kaufmann, "What Is the Offense-Defense Balance and Can We Measure It?," *International Security* 22, no. 4 (1998): 44, https://doi.org/10.2307/2539240.

6. Nadiya Kostyuk and Erik Gartzke, "Why Cyber Dogs Have Yet to Bark Loudly in Russia's Invasion of Ukraine," *Texas National Security Review* 5, no. 3 (Summer 2022): 113–126, https://tnsr.org/2022/06/why-cyber-dogs-have-yet-to-bark-loudly-in-russias-invasion-of-ukraine/. See also Keir Lieber, "The Offense-Defense Balance and Cyber Warfare," in *Cyber Analogies*, eds. Emily O. Goldman and John Arquilla (Monterey, CA: Naval Postgraduate School, 2014), 96–107, https://core.ac.uk/download/pdf/36732393.pdf#page=109.

7. *International Cyberspace and Digital Policy Strategy: Towards an Innovative, Secure, and Rights-Respecting Digital Future* (U.S. Department of State, 2024), 10–14, https://www.state.gov/united-states-international-cyberspace-and-digital-policy-strategy/; *The Cost of Malicious Cyber Activity to the U.S. Economy* (U.S. Council of Economic Advisers, February 16, 2018), https://trumpwhitehouse.archives.gov/wp-content/uploads/2018/03/The-Cost-of-Malicious-Cyber-Activity-to-the-U.S.-Economy.pdf. See also Paul Dreyer, et al., *Estimating the Global Cost of Cyber Risk: Methodology and Examples* (RAND, January 15, 2018), https://www.rand.org/pubs/research_reports/RR2299.html; *Global Cybersecurity Outlook 2024* (World Economic Forum, January 2024), https://www3.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2024.pdf.

8. Grace B. Mueller, et al., *Cyber Operations during the Russo-Ukrainian War: From Strange Patterns to Alternative Futures* (Center for Strategic and International Studies, July 2023), https://www.csis.org/analysis/cyber-operations-during-russo-ukrainian-war; Jon Bateman, Nick Beecroft, and Gavin Wilde, "What the Russian Invasion Reveals About the Future of Cyber Warfare," Carnegie Endowment for International Peace, December 19, 2022, https://carnegieendowment.org/2022/12/19/what-russian-invasion-reveals-about-future-of-cyber-warfare-pub-88667. Regarding the NotPetya attack, see 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/.

9. *Annual Threat Assessment of the U.S. Intelligence Community* (Office of the Director of National Intelligence, February 2023), https://www.dni.gov/files/ODNI/documents/assessments/ATA-2023-Unclassified-Report.pdf.

10. "PRC State-Sponsored Actors Compromise and Maintain Persistent Access to U.S. Critical Infrastructure," Cybersecurity and Infrastructure Security Agency, National Security Agency, and Federal Bureau of Investigation, February 7, 2024, https://www.cisa.gov/news-events/cybersecurity-advisories/aa24-038a.

11. Dakota Cary, *Robot Hacking Games* (Center for Security and Emerging Technology, September 2021), https://cset.georgetown.edu/publication/robot-hacking-games/; John Bansemer, "Soon, the Hackers Won't Be Human," *Foreign Affairs*, December 10, 2021, https://www.foreignaffairs.com/articles/united-states/2021-12-10/soon-hackers-wont-be-human; and Dakota Cary (@DakotaInDC), "Re-upping this paper on destructive cyber attacks on ICS systems enabled by AI. One researcher is tied to Zhejiang Labs, which hosts a cyber range I've tied to China's security services," X, March 1, 2024, https://x.com/DakotaInDC/status/1763585998254952870.

12. Dakota Cary, *Academics, AI, and APTs: How Six Advanced Persistent Threat-Connected Chinese Universities Are Advancing AI Research* (Center for Security and Emerging Technology, March 2021), https://cset.georgetown.edu/publication/academics-ai-and-apts/.

13. Dakota Cary, *China's CyberAI Talent Pipeline* (Center for Security and Emerging Technology, July 2021), https://cset.georgetown.edu/publication/chinas-cyberai-talent-pipeline/.

14. Micah Musser and Ashton Garriott, *Machine Learning and Cybersecurity: Hype and Reality* (Center for Security and Emerging Technology, June 2021), https://cset.georgetown.edu/publication/machine-learning-and-cybersecurity/.

15. *2018 Webroot Threat Report* (Webroot, 2018), https://www-cdn.webroot.com/9315/2354/6488/2018-Webroot-Threat-Report_US-ONLINE.pdf.

16. These dimensions are adapted from Yisroel Mirsky, et al., "The Threat of Offensive AI to Organizations," *Computers & Security* 124 (January 2023): 103006, https://doi.org/10.1016/j.cose.2022.103006.

17. *America's AI Action Plan* (Office of Science and Technology Policy, July 2025), 22, https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf#page=25.41. See also from the Biden administration, "Memorandum on Advancing the United States' Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence," National Security Memorandum 25 (October 24, 2024), https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2024/10/24/memorandum-on-advancing-the-united-states-leadership-in-artificial-intelligence-harnessing-artificial-intelligence-to-fulfill-national-security-objectives-and-fostering-the-safety-security/.

18. *Bipartisan House Task Force Report on Artificial Intelligence* (U.S. House of Representatives, December 2024), https://republicans-science.house.gov/_cache/files/a/a/aa2ee12f-8f0c-46a3-8ff8-8e4215d6a72b/6676530F7A-

30F243A24E254F6858233A.ai-task-force-report-final.pdf; "AI Action Week: List of Side Events," AI Action Summit, Élysée Palace, February 5, 2025, https://www.elysee.fr/admin/upload/default/0001/17/d17d4df370bc061e703ad5346acc24e069d5e9c0.pdf; and "Trust in AI," AI Action Summit, Élysée Palace, January 17, 2025, https://www.elysee.fr/en/sommet-pour-l-action-sur-l-ia/trust-in-ai.

19. Ajeya Cotra and Kelsey Piper, "Language Models Surprised Us," Planned Obsolescence, August 29, 2023, https://www.planned-obsolescence.org/language-models-surprised-us/; Katja Grace, "Survey of 2,778 AI Authors: Six Parts in Pictures," AI Impacts (blog), January 4, 2024, https://blog.aiimpacts.org/p/2023-ai-survey-of-2778-six-things.

20. "Machine Learning Trends," Epoch AI, January 13, 2025, https://epochai.org/trends.

21. "EvalPlus Leaderboard," accessed December 1, 2024, https://evalplus.github.io/leaderboard.html; Zibin Zheng, et al., "A Survey of Large Language Models for Code: Evolution, Benchmarking, and Future Trends," arXiv, November 17, 2023, https://arxiv.org/abs/2311.10372.

22. Nathan Lambert, "GPT-4.5: 'Not a Frontier Model'?" Interconnects, February 28, 2025, https://www.interconnects.ai/p/gpt-45-not-a-frontier-model.

23. Lambert, "GPT-4.5: 'Not a Frontier Model'?"; Jaime Sevilla (@Jsevillamol), "Across models we had observed up until now that a 10x in training compute leads to +10% on GPQA and +20% on MATH. Now we see that 4.5 is 20% better than 4o on GPQA/AIME but people are just not impressed?" X, February 28, 2025, https://x.com/Jsevillamol/status/1895611518672388210.

24. Peter Wildeford, "GPT-5: A Small Step for Intelligence, a Giant Leap for Normal People," The Power Law (blog), August 8, 2025, https://peterwildeford.substack.com/p/gpt-5-a-small-step-for-intelligence; Epoch AI (@EpochAIResearch), "OpenAI has historically scaled up training compute by around 100x with each new generation of its GPT. However, GPT-5 appears to be an exception to this trend." X, August 8, 2025, https://x.com/EpochAIResearch/status/1953883611389702169.

25. Ege Erdil, "What AI Can Currently Do Is Not the Story," Epoch AI, March 7, 2025, https://epoch.ai/gradient-updates/what-ai-can-currently-do-is-not-the-story. For more recent analysis of the pace of AI progress through the release of GPT-5, see Wildeford, "GPT-5: A Small Step for Intelligence, a Giant Leap for Normal People."

26. On reasoning behaviors as a baseline to strengthen, see Kanishk Gandhi, et al., "Cognitive Behaviors That Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs," arXiv, March 3, 2025, https://arxiv.org/abs/2503.01307; DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," arXiv, January 22, 2025, sec. 4.1, https://arxiv.org/abs/2501.12948.

27. Jaime Sevilla, et al., *Can AI Scaling Continue Through 2030?* (Epoch AI, August 20, 2024), https://epochai.org/blog/can-ai-scaling-continue-through-2030; Dwarkesh Patel, "Will Scaling Work?" Dwarkesh Podcast (blog), December 26, 2023, https://www.dwarkeshpatel.com/p/will-scaling-work; Arvind Narayanan and Sayash Kapoor, "AI Scaling Myths," AI Snake Oil, June 27, 2024, https://www.aisnakeoil.com/p/ai-scaling-myths; and Ethan Mollick, "Scaling: The State of Play in AI," One Useful Thing, September 16, 2024, https://www.oneusefulthing.org/p/scaling-the-state-of-play-in-ai. The view that deep learning is "hitting a wall" has been advanced by Gary Marcus in particular (Gary Marcus, "Deep Learning Is Hitting a Wall," Nautilus, March 10, 2022, https://nautil.us/deep-learning-is-hitting-a-wall-238440/).

28. Bengio, et al., *International AI Safety Report*, sec. 2.1.3.

29. On the limitations of frontier AI, see Bengio, et al., *International AI Safety Report*, 43, 51–54, 75.

30. For relevant discussion and caveats, see Andrew Lohn, et al., *Autonomous Cyber Defense: A Roadmap from Lab to Ops* (Center for Security and Emerging Technology, June 2023), https://cset.georgetown.edu/publication/autonomous-cyber-defense/; Ben Buchanan, et al., *Automating Cyber Attacks: Hype and Reality* (Center for Security and Emerging Technology, November 2020), https://cset.georgetown.edu/publication/automating-cyber-attacks/.

31. "Re: Request for Information (RFI) on the Development of an Artificial Intelligence (AI) Action Plan ("Plan")," Anthropic, March 6, 2025, https://assets.anthropic.com/m/4e20a4ab6512e217/original/Anthropic-Response-to-OSTP-RFI-March-2025-Final-Submission-v3.pdf.

32. Sergei Glazunov and Mark Brand, "Project Naptime: Evaluating Offensive Security Capabilities of Large Language Models," Google Project Zero, June 20, 2024, https://googleprojectzero.blogspot.com/2024/06/project-naptime.html; Yuxuan Zhu, et al., "Teams of LLM Agents Can Exploit Zero-Day Vulnerabilities," arXiv, June 2, 2024, https://arxiv.org/abs/2406.01637; Jiacen Xu, et al., "AutoAttacker: A Large Language Model Guided System to Implement Automatic Cyber-attacks," arXiv, March 2, 2024, https://arxiv.org/abs/2403.01038; Andreas Happe, Aaron Kaplan, and Juergen Cito, "LLMs as Hackers: Autonomous Linux Privilege Escalation Attacks," arXiv, August 1, 2024, https://arxiv.org/abs/2310.11409; Richard Fang, et al., "LLM Agents Can Autonomously Exploit One-day Vulnerabilities," arXiv, April 17, 2024, https://arxiv.org/abs/2404.08144; and Richard Fang, et al., "LLM Agents Can Autonomously Hack Websites," arXiv, February 16, 2024, https://arxiv.org/abs/2402.06664. Note that the last two papers have been critiqued as overstating the level of capability demonstrated—see Chris Rohlf, "No, LLM Agents Can Not Autonomously Exploit One-day Vulnerabilities," Root Cause, April 21, 2024, https://struct.github.io/auto_agents_1_day.html; Chris Rohlf, "No, LLM Agents Cannot Autonomously "Hack" Websites," Root Cause, February 19, 2024, https://struct.github.io/llm_auto_hax.html. They are nonetheless included as examples where a model—in this case, GPT-4—substantially outperformed previously available models in a cybersecurity-related capability evaluation.

33. *System Card: Claude Opus 4 & Claude Sonnet 4* (Anthropic, May 2025), sec. 7.4, https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf; Andy K. Zhang, et al., "Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models," in *The Thirteenth International Conference on Learning Representations* (2025), https://cybench.github.io/.

34. *OpenAI o3 and o4-mini System Card* (OpenAI, April 16, 2025), sec. 4.3, https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf. Some caveats apply to these figures—for example, they are

based on the best 12 out 16 attempts; and performance was even higher when models had access to the internet (but this may reflect that hints specific to some of the evaluated capture-the-flag challenges were available online).

35. *GPT-5 System Card*, OpenAI, sec. 5.1.2, https://cdn.openai.com/pdf/8124a3ce-ab78-4f06-96eb-49ea29ffb52f/gpt5-system-card-aug7.pdf#page=27.82.

36. Oege de Moor and Albert Ziegler, "XBOW Unleashes GPT-5's Hidden Hacking Power, Doubling Performance," XBOW (blog), August 15, 2025, https://xbow.com/blog/gpt-5; Oege de Moor, "XBOW Now Matches the Capabilities of a Top Human Pentester," XBOW (blog), August 5, 2024, https://xbow.com/blog/xbow-vs-humans; and Nico Waisman, "XBOW Validation Benchmarks: Show Me the Numbers!" XBOW (blog), September 11, 2024, https://xbow.com/blog/benchmarks/. The XBOW Challenge featured an original set of capture-the-flag challenges sourced from penetration testing firms. Human penetration testers were given 40 hours to solve as many as possible out of 104 challenges; the AI system completed its attempt in 28 minutes. A "principal" penetration tester with two decades of cybersecurity experience was still able to match the performance of the XBOW system, including solving more "hard"-difficulty problems.

37. Robi Rahman, et al., "Over 20 AI Models Have Been Trained at the Scale of GPT-4," Epoch AI, January 30, 2025, https://epoch.ai/data-insights/models-over-1e25-flop.

38. Lorenzo Franceschi-Bicchierai, "Google Says Its AI-Based Bug Hunter Found 20 Security Vulnerabilities," TechCrunch, August 4, 2025, https://techcrunch.com/2025/08/04/google-says-its-ai-based-bug-hunter-found-20-security-vulnerabilities/; "Final Competition Winners Announcement," AI Cyber Challenge, August 2025; XBOW, "Blog," accessed December 20, 2024, https://xbow.com/blog/; Kent Walker, "A Summer of Security: Empowering Cyber Defenders with AI," Google, July 15, 2025, https://blog.google/technology/safety-security/cybersecurity-updates-summer-2025/; Sean Heelan, "How I Used o3 to Find CVE-2025-37899, a Remote Zeroday Vulnerability in the Linux Kernel's SMB Implementation," May 22, 2025, https://sean.heelan.io/2025/05/22/how-i-used-o3-to-find-cve-2025-37899-a-remote-zeroday-vulnerability-in-the-linux-kernels-smb-implementation/; and "Argusee: A Multi-Agent Collaborative Architecture for Automated Vulnerability Discovery," DARKNAVY, May 23, 2025, https://www.darknavy.org/blog/argusee_a_multi_agent_collaborative_architecture_for_automated_vulnerability_discovery/.

39. *The Near-Term Impact of AI on the Cyber Threat* (National Cyber Security Centre (United Kingdom), January 24, 2024), https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat; *Secure, Empower, Advance: How AI Can Reverse the Defender's Dilemma* (Google, February 2024), 7, https://services.google.com/fh/files/misc/how-ai-can-reverse-defenders-dilemma.pdf. See also Maia Hamin and Stewart Scott, *Hacking with AI: The Use of Generative AI in Malicious Cyber Activity* (Atlantic Council, February 2024), https://dfrlab.org/wp-content/uploads/sites/3/2024/02/csi-report-hacking-with-ai.pdf.

40. *Leveraging AI to Enhance American Communications: Hearing before the Subcommittee on Communications and Technology of the House Committee on Energy and Commerce*, 118th Cong. (2023) (testimony of Sam Rubin, Vice President – Global Operations, Unit 42, Palo Alto Networks), https://d1dth6e84htgma.cloudfront.net/11_14_23_Rubin_Testimony_2fba2978dd.pdf.

41. Jennifer Tang, Tiffany Saade, and Steve Kelly, *The Implications of Artificial Intelligence in Cybersecurity: Shifting the Offense-Defense Balance* (Institute for Security and Technology, October 2024), 11–13, https://securityandtechnology.org/wp-content/uploads/2024/10/The-Implications-of-Artificial-Intelligence-in-Cybersecurity.pdf.

42. "Disrupting Malicious Uses of AI by State-Affiliated Threat Actors," OpenAI, February 14, 2024, https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/; "Staying Ahead of Threat Actors in the Age of AI," Microsoft Threat Intelligence (blog), February 14, 2024, https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/; and Google Threat Intelligence Group, "Adversarial Misuse of Generative AI," Threat Intelligence (blog), January 29, 2025, https://cloud.google.com/blog/topics/threat-intelligence/adversarial-misuse-generative-ai.

43. Vitaly Simonovich, "Cato CTRL™ Threat Research: Analyzing LAMEHUG – First Known LLM-Powered Malware with Links to APT28 (Fancy Bear)," Cato Networks, July 23, 2025, https://www.catonetworks.com/blog/cato-ctrl-threat-research-analyzing-lamehug/; Alex Delamotte, "Predator AI | ChatGPT-Powered Infostealer Takes Aim at Cloud Platforms," SentinelOne, November 7, 2023, https://www.sentinelone.com/labs/predator-ai-chatgpt-powered-infostealer-takes-aim-at-cloud-platforms/; "BlackMamba: AI-Synthesized, Polymorphic Keylogger with On-the-Fly Program Modification," HYAS, 2023, https://www.hyas.com/hubfs/Downloadable%20Content/HYAS-AI-Augmented-Cyber-Attack-WP-1.1.pdf; and Bernhard Mueller (@muellerberndt), "This is so cool and scary at the same time! I was able to run DarwinGPT with #GPT4 and it quickly evolved into a passable "worm" ...," X, April 11, 2023, https://x.com/muellerberndt/status/1645651344773222400.

44. Timothée Chauvin, "24 Theses on Cybersecurity," October 5, 2024, thesis 2, https://tchauvin.com/theses-on-cybersecurity-and-ai.

45. "Microsoft Security Copilot," Microsoft, https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-copilot-security; "Purple AI," SentinelOne, https://www.sentinelone.com/platform/purple/; and "Charlotte AI," CrowdStrike, https://www.crowdstrike.com/platform/charlotte-ai/.

46. Ben Edelman, et al., *Randomized Controlled Trial for Copilot for Security: Whitepaper* (Microsoft, January 2024), https://www.microsoft.com/content/dam/microsoft/final/en-us/microsoft-product-and-services/microsoft-dynamics-365/pdf/Microsoft-Copilot-for-Security-productivity-findings-Whitepaper-Jan2024.pdf.

47. Sydney J. Freedberg Jr., "'I'm Disappointed': Pentagon CIO Cybersecurity Chief Asks Industry, Where's My AI?," Breaking Defense, March 13, 2024, https://breakingdefense.com/2024/03/im-disappointed-pentagon-cio-cybersecurity-chief-asks-industry-wheres-my-ai/.

48. "Pilot for Artificial Intelligence Enabled Vulnerability Detection," Cybersecurity and Infrastructure Security Agency, July 29, 2024, https://www.cisa.gov/resources-tools/resources/pilot-artificial-intelligence-enabled-vulnerability-detection.

49. Andrew Lohn and Krystal Jackson, *Will AI Make Cyber Swords or Shields?* (Center for Security and Emerging

Technology, August 2022), https://cset.georgetown.edu/publication/will-ai-make-cyber-swords-or-shields/; Bruce Schneier, "Should U.S. Hackers Fix Cybersecurity Holes or Exploit Them?," *The Atlantic*, May 19, 2014, https://www.theatlantic.com/technology/archive/2014/05/should-hackers-fix-cybersecurity-holes-or-exploit-them/371197/; Jonathan M. Spring, "An Analysis of How Many Undiscovered Vulnerabilities Remain in Information Systems," *Computers & Security* 131 (August 2023): 103191, https://doi.org/10.1016/j.cose.2023.103191; Mingyi Zhao and Peng Liu, "Empirical Analysis and Modeling of Black-Box Mutational Fuzzing," in *Engineering Secure Software and Systems*, eds. Juan Caballero, Eric Bodden, and Elias Athanasopoulos (Cham: Springer International Publishing, 2016), 173–189, https://doi.org/10.1007/978-3-319-30806-7_11; and O.H. Alhazmi, Y.K. Malaiya, and I. Ray, "Measuring, Analyzing and Predicting Security Vulnerabilities in Software Systems," *Computers & Security* 26, no. 3 (May 2007): 219–228, https://doi.org/10.1016/j.cose.2006.10.002. For a critical perspective on the extent that cyberspace is offense-dominant, see Lieber, "The Offense-Defense Balance and Cyber Warfare"; Rebecca Slayton, "What Is the Cyber Offense-Defense Balance? Conceptions, Causes, and Assessment," *International Security* 41, no. 3 (Winter 2016/17): 72–109, https://direct.mit.edu/isec/article-abstract/41/3/72/12149/What-Is-the-Cyber-Offense-Defense-Balance.

50. Ben Buchanan, *The Cybersecurity Dilemma: Hacking, Trust, and Fear between Nations* (Oxford University Press, 2017).

51. U.S. Cyber Command Public Affairs Office, "CYBER 101 - Defend Forward and Persistent Engagement," October 25, 2022, https://www.cybercom.mil/Media/News/Article/3198878/cyber-101-defend-forward-and-persistent-engagement/.

52. See also Jacquelyn Schneider, "The Information Revolution and Offense-Defense Balance in U.S. Doctrine," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, November 20, 2021), https://doi.org/10.2139/ssrn.3967772; Jacquelyn Schneider, "The Capability/Vulnerability Paradox and Military Revolutions: Implications for Computing, Cyber, and the Onset of War," *Journal of Strategic Studies* 42, no. 6 (September 19, 2019): 841–863, https://doi.org/10.1080/01402390.2019.1627209; and Cullen O'Keefe, "Widespread Access to Defense-Dominant Technologies Can Still Increase Risk," Jural Networks, May 17, 2024, https://juralnetworks.substack.com/p/widespread-access-to-defense-dominant.

53. Timothée Chauvin (@timotheechauvin), "In fact, I believe that AI vuln detection in source code will favor cyberdefense much more than fuzzing did ...," X, April 26, 2024, https://x.com/timotheechauvin/status/1783785367885279683.

54. Spring, "An Analysis of How Many Undiscovered Vulnerabilities Remain in Information Systems."

55. Karina Nguyen, "The Cost of AI Reasoning Is Going to Drastically Decrease," sémaphore, March 29, 2024, https://semaphore.substack.com/p/the-cost-of-reasoning-in-raw-intelligence; Guido Appenzeller, "Welcome to LLMflation – LLM Inference Cost Is Going Down Fast ," Andreessen Horowitz, November 12, 2024, https://a16z.com/llmflation-llm-inference-cost/; Artificial Analysis, "AI Review 2024 Highlights," 6, https://artificialanalysis.ai/downloads/ai-review/2024/Artificial-Analysis-AI-Review-2024-Highlights.pdf; Ben Cottier, et al., "LLM Inference Prices Have Fallen Rapidly but Unequally Across Tasks," Epoch AI, 2025, https://epoch.ai/data-insights/llm-inference-price-trends; and Harlan Lewis, "LLM Capability, Cost, & Throughput," https://docs.google.com/spreadsheets/d/1foc98Jtbi0-GUsNySd-dvL0b2a7EuVQw8MoaQIWaDT-w/edit?usp=sharing.

56. Toby Ord, "Inference Scaling and the Log-x Chart," January 21, 2025, https://www.tobyord.com/writing/inference-scaling-and-the-log-x-chart.

57. "Details about METR's Preliminary Evaluation of o3 and o4-mini," METR, April 16, 2025, https://metr.github.io/autonomy-evals-guide/openai-o3-report/.

58. Zachary Fryer-Biggs, "Secretive Pentagon Research Program Looks to Replace Human Hackers with AI," Yahoo News, September 13, 2020, https://www.yahoo.com/news/secretive-pentagon-research-program-looks-to-replace-human-hackers-with-ai-090032920.html.

59. Kostyuk and Gartzke, "Why Cyber Dogs Have Yet to Bark Loudly in Russia's Invasion of Ukraine." See also Lieber, "The Offense-Defense Balance and Cyber Warfare." On further factors explaining the modest role of cyberattacks in Russia's invasion of Ukraine, see Bateman, Beecroft, and Wilde, "What the Russian Invasion Reveals about the Future of Cyber Warfare"; Dominika Dziwisz and Błażej Sajduk, "The Russia-Ukraine Conflict from 2014 to 2023 and the Significance of a Strategic Victory in Cyberspace," *Applied Cybersecurity & Internet Governance* 2, no. 1 (2023): 1–20, https://doi.org/10.60097/ACIG/162842; and Thomas Macaulay, "The War in Ukraine Is Exposing the Limits of Cyber Warfare — and Russian Hackers," TNW, November 24, 2022, https://thenextweb.com/news/why-russia-cyber-army-has-struggled-to-impact-ukraine-war.

60. Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (W. W. Norton & Company, 2018), chap. 14.

61. For analysis of frontier AI strengths and weaknesses across the end-to-end attack chain, see Mikel Rodriguez, et al., "A Framework for Evaluating Emerging Cyberattack Capabilities of AI," arXiv, March 14, 2025, https://arxiv.org/abs/2503.11917.

62. Andy Greenberg, "How an Entire Nation Became Russia's Test Lab for Cyberwar," *Wired*, June 20, 2017, https://www.wired.com/story/russian-hackers-attack-ukraine/; "CrashOverride Malware," Cybersecurity and Infrastructure Security Agency, June 12, 2017, https://www.cisa.gov/news-events/alerts/2017/06/12/crashoverride-malware; Microsoft Threat Intelligence, "Volt Typhoon Targets US Critical Infrastructure with Living-Off-the-Land Techniques," Microsoft Security Blog, May 24, 2023, https://www.microsoft.com/en-us/security/blog/2023/05/24/volt-typhoon-targets-us-critical-infrastructure-with-living-off-the-land-techniques/; and Cybersecurity and Infrastructure Security Agency, National Security Agency, and Federal Bureau of Investigation, "PRC State-Sponsored Actors Compromise and Maintain Persistent Access to U.S. Critical Infrastructure."

63. "Network Traffic, Data Source DS0029," MITRE ATT&CK, https://attack.mitre.org/datasources/DS0029/.

64. Kim Zetter, "The Untold Story of the Boldest Supply-Chain Hack Ever," *Wired*, May 2, 2023, https://www.wired.com/story/the-untold-story-of-solarwinds-the-boldest-supply-chain-hack-ever/.

65. Ellen Nakashima and Joby Warrick, "Stuxnet Was Work of U.S. and Israeli Experts, Officials Say," *The Washington Post*, June 2, 2012, https://www.washingtonpost.com/world/national-security/stuxnet-was-work-of-us-and-israeli-experts-officials-say/2012/06/01/gJQAlnEy6U_story.html.

66. Nathan Benaich and Alex Chalmers, "Open-Endedness Is All We'll Need: On "Agentic AI,"" Air Street Press, August 15, 2024, https://press.airstreet.com/p/open-endedness-is-all-well-need; "Details about METR's Preliminary Evaluation of OpenAI o1-Preview," METR, September 2024, https://metr.github.io/autonomy-evals-guide/openai-o1-preview-report/; and Anthropic, "Introducing Computer Use, a New Claude 3.5 Sonnet, and Claude 3.5 Haiku," October 22, 2024, https://www.anthropic.com/news/3-5-models-and-computer-use.

67. Bengio, et al., *International AI Safety Report*, 43, 51–54, 75; Shengye Wan, et al., "CYBERSECEVAL 3: Advancing the Evaluation of Cybersecurity Risks and Capabilities in Large Language Models," arXiv, August 2, 2024, sec. 3.3–3.4, https://arxiv.org/abs/2408.01605; and "Details about METR's Preliminary Evaluation of OpenAI o1-Preview."

68. Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati, "LLMs Still Can't Plan; Can LRMs? A Preliminary Evaluation of OpenAI's o1 on PlanBench," in NeurIPS 2024 Workshop on Open-World Agents, 2024, https://openreview.net/forum?id=Gcr1Lx4Koz.

69. Thomas Kwa, et al., "Measuring AI Ability to Complete Long Tasks," arXiv, March 18, 2025, https://arxiv.org/abs/2503.14499.

70. Oriol Vinyals, et al., "Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning," *Nature* 575 (2019): 350–354, https://www.nature.com/articles/s41586-019-1724-z; Christopher Berner, et al., "Dota 2 with Large Scale Deep Reinforcement Learning," arXiv, December 13, 2019, https://arxiv.org/abs/1912.06680.

71. OpenAI, "Learning to Reason with LLMs," September 12, 2024, https://openai.com/index/learning-to-reason-with-llms/; DeepSeek-AI, "DeepSeek-R1."

72. Shane Caldwell, et al., "PentestJudge: Judging Agent Behavior Against Operational Requirements," arXiv, August 4, 2025, https://arxiv.org/abs/2508.02921.

73. Anthropic, "Developing a Computer Use Model," October 22, 2024, https://www.anthropic.com/news/developing-computer-use.

74. Anne Johnson and Emily Grumbling, Rapporteurs, *Implications of Artificial Intelligence for Cybersecurity: Proceedings of a Workshop* (The National Academies Press, 2019), 13–14, https://nap.nationalacademies.org/catalog/25488/implications-of-artificial-intelligence-for-cybersecurity-proceedings-of-a-workshop; Paul Scharre, *Four Battlegrounds: Power in the Age of Artificial Intelligence* (W. W. Norton & Company, 2024), chap. 24.

75. Tom Davidson, et al., "AI Capabilities Can Be Significantly Improved without Expensive Retraining," arXiv, December 12, 2023, https://doi.org/10.48550/arXiv.2312.07413; Usman Anwar, et al., "Foundational Challenges in Assuring Alignment and Safety of Large Language Models," Transactions on Machine Learning Research, September 2, 2024, sec. 2.2, https://openreview.net/forum?id=oVTkOs8Pka.

76. Glazunov and Brand, "Project Naptime."

77. Anwar, et al., "Foundational Challenges in Assuring Alignment and Safety of Large Language Models," sec. 3.2 and 3.5; Adam Gleave, "AI Safety in a World of Vulnerable Machine Learning Systems," FAR.AI, October 7, 2024, https://www.alignmentforum.org/posts/ncsxcf8CkDveXBCrA/ai-safety-in-a-world-of-vulnerable-machine-learning-systems-1.

78. Abhay Sheshadri, et al., "Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs," OpenReview, September 25, 2024, https://openreview.net/forum?id=wI5uHZLeCZ.

79. Kylie Robison, "OpenAI Teases New Reasoning Model—but Don't Expect to Try It Soon," The Verge, December 20, 2024, https://www.theverge.com/2024/12/20/24326036/openai-o1-o2-o3-reasoning-model-testing.

80. Anwar, et al., "Foundational Challenges in Assuring Alignment and Safety of Large Language Models," sec. 3.5.

81. Bansemer, "Soon, the Hackers Won't Be Human."

82. Kim Zetter, "How Digital Detectives Deciphered Stuxnet, the Most Menacing Malware in History," *Wired*, July 11, 2011, https://www.wired.com/2011/07/how-digital-detectives-deciphered-stuxnet/.

83. Greenberg, "The Untold Story of NotPetya."

84. Apostol Vassilev, et al., "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations" (National Institute of Standards and Technology, January 2024), https://csrc.nist.gov/pubs/ai/100/2/e2023/final.

85. gwern, June 9, 2024, Re: Buck Shlegeris, "Access to Powerful AI Might Make Computer Security Radically Easier," AI Alignment Forum, June 8, 2024, https://www.lesswrong.com/posts/2wxufQWK8rXcDGbyL/access-to-powerful-ai-might-make-computer-security-radically?commentId=JEWjT2ByniSq95QDj#JEWjT2ByniSq95QDj.

86. Anwar, et al., "Foundational Challenges in Assuring Alignment and Safety of Large Language Models," sec. 3.1 and 3.2.

87. Jan Betley, et al., "Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs," arXiv, February 24, 2025, https://arxiv.org/abs/2502.17424.

88. Ryan Greenblatt, et al., "Alignment Faking in Large Language Models," arXiv, December 18, 2024, https://arxiv.org/abs/2412.14093; Alexander Meinke, et al., "Frontier Models are Capable of In-Context Scheming," arXiv, December 6, 2024, https://arxiv.org/abs/2412.04984.

89. Timothée Chauvin (@timotheechauvin), "Front-running security patches in the age of fast reverse-engineering agents: ...," X, July 19, 2024, https://x.com/timotheechauvin/status/1814242466607820995.

90. Casey Charrier and Robert Weiner, "How Low Can You Go? An Analysis of 2023 Time-to-Exploit Trends," Mandiant, October 15, 2024, https://cloud.google.com/blog/topics/threat-intelligence/time-to-exploit-trends-2023.

91. Jessica Lyons, "CrowdStrike Apologizes to Congress for 'Perfect Storm' That Caused Global IT Outage," The Register, September 25, 2024, https://www.theregister.com/2024/09/25/crowdstrike_to_congress_perfect_storm.

92. Drexel and Withers, *Catalyzing Crisis,* sec. "Integrating AI into Complex Systems."

93. Bengio, et al., *International AI Safety Report*, sec. 2.3.3.

94. Jacquelyn Schneider, *Digitally-Enabled Warfare: The Capability-Vulnerability Paradox* (Center for a New American Security, August 2016), https://www.cnas.org/publications/reports/digitally-enabled-warfare-the-capability-vulnerability-paradox; Schneider, "The Capability/Vulnerability Paradox and Military Revolutions."

95. Google DeepMind, "Frontier Safety Framework, Version 2.0," February 4, 2025, 1, https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/updating-the-frontier-safety-framework/Frontier%20Safety%20Framework%202.0%20(1).pdf; Anthropic, "Responsible Scaling Policy: Effective October 15, 2024," October 15, 2024, fn. 18, https://assets.anthropic.com/m/24a47b00f10301cd/original/Anthropic-Responsible-Scaling-Policy-2024-10-15.pdf.

96. See, for example, OpenAI, "Hello GPT-4o," May 13, 2024, https://openai.com/index/hello-gpt-4o/; Gemini Team (Google), "Gemini: A Family of Highly Capable Multimodal Models," arXiv, December 19, 2023, https://doi.org/10.48550/arXiv.2312.11805.

97. "Details about METR's Preliminary Evaluation of OpenAI o1-Preview."

98. Jeremy Kahn and Beatrice Nolan, "Google Released Safety Risks Report of Gemini 2.5 Pro Weeks after Its Release — but an AI Governance Expert Said It Was a 'Meager' and 'Worrisome' Report," *Fortune*, April 17, 2025, https://fortune.com/article/google-gemini-2-5-pro-model-card-published-ai-governance-expert-criticizes-it-as-meager-and-worrisome/.

99. OpenAI, "Introducing Deep Research," February 2, 2025, https://openai.com/index/introducing-deep-research/; *Deep Research System Card* (Open AI, February 25, 2025), https://cdn.openai.com/deep-research-system-card.pdf.

100. Jeremy Kahn, "Top AI Labs Aren't Doing Enough to Ensure AI Is Safe, a Flurry of Recent Datapoints Suggest," *Fortune*, December 17, 2024, https://fortune.com/2024/12/17/openai-o1-deception-unsafe-safety-testing-future-of-life-institute-grades/.

101. Beatrice Nolan, "Elon Musk Released XAI's Grok 4 Without Any Safety Reports—Despite Calling AI More 'Dangerous Than Nukes,'" *Fortune*, July 17, 2025, https://fortune.com/2025/07/17/elon-musk-xai-grok-4-no-safety-report/; *Grok 4 Model Card* (xAI, August 20, 2025), https://data.x.ai/2025-08-20-grok-4-model-card.pdf.

102. Billy Perrigo, "Inside the U.K.'s Bold Experiment in AI Safety," *Time*, January 16, 2025, https://time.com/7204670/uk-ai-safety-institute/; "U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing and Evaluation with Anthropic and OpenAI," National Institute of Standards and Technology, August 29, 2024, https://www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research; and Maria Curi and Ashley Gold, "Elon Musk Deal With AI Safety Office Persists Amid DOGE Scrutiny," Axios Pro, March 4, 2025, https://www.axios.com/pro/tech-policy/2025/03/04/elon-musk-deal-with-ai-safety-office-persists-amid-doge-scrutiny.

103. Danny Tobey, et al., "China Releases AI Safety Governance Framework," DLA Piper, September 12, 2024, https://www.dlapiper.com/en-us/insights/publications/2024/09/china-releases-ai-safety-governance-framework.

104. O'Keefe, "Widespread Access to Defense-Dominant Technologies Can Still Increase Risk."

105. Davidson, et al., "AI Capabilities Can Be Significantly Improved without Expensive Retraining."

106. Anthropic, "Responsible Scaling Policy," September 19, 2023, https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf, as cited in Caleb Withers, "Response to NTIA Request for Comment: "Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights,"" Center for a New American Security, 13, https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS-NTIA-Open-Weights-response.pdf.

107. See also Wyatt Hoffman, *AI and the Future of Cyber Competition* (Center for Security and Emerging Technology, January 2021), https://cset.georgetown.edu/publication/ai-and-the-future-of-cyber-competition/.

108. Greenblatt, et al., "Alignment Faking in Large Language Models"; Meinke, et al., "Frontier Models are Capable of In-Context Scheming."

109. Bengio, et al., *International AI Safety Report*, sec. 2.2.3. On anticipating and preventing self-exfiltration, see also Hjalmar Wijk, "Autonomous Replication and Adaptation: An Attempt at a Concrete Danger Threshold," AI Alignment Forum, August 16, 2023, https://www.alignmentforum.org/posts/vERGLBpDE8m5mpT6t/autonomous-replication-and-adaptation-an-attempt-at-a; Jan Leike, "Self-Exfiltration Is a Key Dangerous Capability," Musings on the Alignment Problem, September 13, 2023, https://aligned.substack.com/p/self-exfiltration.

110. See "Secure by Design Pledge," Cybersecurity and Infrastructure Security Agency, 2024, https://www.cisa.gov/securebydesign/pledge; Christoph Kern, "Secure by Design at Google," Google, March 4, 2024, https://storage.googleapis.com/gweb-research2023-media/pubtools/7661.pdf.

111. Bob Lord, Jack Cable, and Lauren Zabierek, "Categorically Unsafe Software: Time to Class Up the Joint!" Cybersecurity and Infrastructure Security Agency, May 13, 2024, https://www.cisa.gov/news-events/news/categorically-unsafe-software.

112. Jiwon Ma and Mark Montgomery, *2024 Annual Report on Implementation* (CSC 2.0, September 2024), https://cybersolarium.org/annual-assessment/2024-annual-report-on-implementation/; see also Foundation for Defense of Democracies, "America's Cyber Resiliency in 2024: A Conversation with CSC 2.0 Co-Chair Sen. Angus King," video, 55:35, streamed live on September 19, 2024, https://www.youtube.com/watch?v=3K2lsJsJyGU.

113. See also *Cybersecurity: National Cyber Director Needs to Take Additional Actions to Implement an Effective Strategy*, GAO-24-106916 (U.S. Government Accountability Office, February 1, 2024), https://www.gao.gov/products/gao-24-106916; Joseph Menn, "Alarmed by Chinese Hacks, Republicans Mute Attacks on Cybersecurity Agency," *The Washington Post*, February 3, 2025, https://www.washingtonpost.

com/technology/2025/02/03/cisa-china-trump-no-em-hacking-cyberthreats/; and Tim Starks, "House Appropriators Have Reservations — or Worse — About Proposed CISA Cuts," CyberScoop, May 6, 2025, https://cyberscoop.com/house-questions-trump-cisa-budget-cuts-2025/.

114. See also *Regulatory Harm or Harmonization? Examining the Opportunity to Improve the Cyber Regulatory Regime: Hearing Before the House Committee on Homeland Security*, 118th Cong. (March 11, 2025), https://homeland.house.gov/hearing/regulatory-harm-or-harmonization-examining-the-opportunity-to-improve-the-cyber-regulatory-regime/.

115. *America's AI Action Plan*, 9–10.

116. Bengio, et al., *International AI Safety Report*, sec. 3.4.3; Miles Brundage, et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," arXiv, February 2018, 96, https://arxiv.org/abs/1802.07228; and Fast-Track Action Committee on Advancing Privacy-Preserving Data Sharing and Analytics, Networking and Information Technology Research and Development Subcommittee of the National Science and Technology Council, *National Strategy to Advance Privacy-Preserving Data Sharing and Analytics* (National Coordination Office, Networking and Information Technology Research and Development Program, March 2023), https://www.nitrd.gov/pubs/National-Strategy-to-Advance-Privacy-Preserving-Data-Sharing-and-Analytics.pdf.

117. See also Spring, "An Analysis of How Many Undiscovered Vulnerabilities Remain in Information Systems."

118. Tang, Saade, and Kelly, *The Implications of Artificial Intelligence in Cybersecurity*, rec. C; Alan Chan, et al., "IDs for AI Systems," arXiv, October 28, 2024, https://arxiv.org/abs/2406.12137.

119. Herbie Bradley and Girish Sastry, "The Great Refactor: How to Secure Critical Open-Source Code Against Memory Safety Exploits by Automating Code Hardening at Scale," Institute for Progress, August 11, 2025, https://ifp.org/the-great-refactor/; Chris Rohlf, "AI and the Software Vulnerability Lifecycle," Center for Security and Emerging Technology, August 4, 2025, https://cset.georgetown.edu/article/ai-and-the-software-vulnerability-lifecycle/; and Clark Barrett, et al., "Identifying and Mitigating the Security Risks of Generative AI," *Foundations and Trends® in Privacy and Security* 6, no. 1 (December 14, 2023): 1–52, sec. 5, https://doi.org/10.1561/3300000041.

120. For further discussion on the value of third-party evaluations, see *Draft Report of the Joint California Policy Working Group on AI Frontier Models* (Joint California Policy Working Group on AI Frontier Models, March 2025), sec. 3.2, https://www.cafrontieraigov.org/wp-content/uploads/2025/03/Draft_Report_of_the_Joint_California_Policy_Working_Group_on_AI_Frontier_Models.pdf; *Third-Party Assessments* (Frontier Model Forum, August 4, 2025), https://www.frontiermodelforum.org/technical-reports/third-party-assessments/.

121. "Memorandum on Advancing the United States' Leadership in Artificial Intelligence."

122. *America's AI Action Plan,* 22.

123. See, for example, relevant portions of the Future of AI Innovation Act (U.S. Senate Commerce, Science and Transportation Committee, "Commerce Committee Passes Bipartisan Bill to Ensure U.S. Leads Global AI Innovation," press release, July 31, 2024, https://www.commerce.senate.gov/2024/7/commerce-committee-passes-bipartisan-bill-to-ensure-u-s-leads-global-ai-innovation) and AI Advancement and Reliability Act, H.R. 9497, 118th Cong. (2024), https://science.house.gov/2024/9/h-r-xxxx-ai-advancement-and-reliability-act.

124. "Commerce Proposes Reporting Requirements for Frontier AI Developers and Compute Providers," Bureau of Industry and Security, September 9, 2024, https://www.bis.gov/press-release/commerce-proposes-reporting-requirements-frontier-ai-developers-compute-providers.

125. Ashley Mehra, "Executive Orders on AI: How to (Lawfully) Apply the Defense Production Act," Mercatus Center, January 21, 2025, https://www.mercatus.org/research/policy-briefs/executive-orders-ai-how-lawfully-apply-defense-production-act; "AG Reyes Leads 20-State Letter Asking AI to be Regulated by the People, Not Politics," Office of the Utah Attorney General, February 5, 2024, https://attorneygeneral.utah.gov/2024/02/05/ag-reyes-leads-20-state-letter-asking-ai-to-be-regulated-by-the-people-not-politics/.

126. On the one hand, continued algorithmic progress will enable increasingly powerful models to be trained with a given amount of compute, suggestive of a need to lower thresholds over time. On the other hand, it may be justifiable to raise thresholds over time: as regulators gather evidence about the societal impacts of increasingly powerful models; as relevant mitigations are implemented; and as increasingly affordable compute ultimately necessitates pragmatism. See *Draft Report of the Joint California Policy Working Group on AI Frontier Models*, sec. 5.2; Anson Ho, et al., "Algorithmic Progress in Language Models," Epoch AI, March 12, 2024, https://epochai.org/blog/algorithmic-progress-in-language-models; Lennart Heim and Leonie Koessler, "Training Compute Thresholds: Features and Functions in AI Regulation," arXiv, August 6, 2024, https://arxiv.org/abs/2405.10799; and Paul Scharre, "Future-Proofing Frontier AI Regulation" (Center for a New American Security, March 2024), 34, https://www.cnas.org/publications/reports/future-proofing-frontier-ai-regulation.

127. See also Mahmoud Ghanem, "Cyber Range Design at UK AISI" (talk presented at AI Security Forum, Paris, 2025), https://far.ai/events/sessions/mahmoud-ghanem-cyber-range-design-at-uk-aisi.

128. Hamin and Scott, *Hacking with AI*, 19.

129. For (inference) cost and time taken, respectively, see, for example, Zhu, et al., "Teams of LLM Agents Can Exploit Zero-Day Vulnerabilities." On best practices for human baselines, see Kevin L. Wei, et al., "Recommendations and Reporting Checklist for Rigorous & Transparent Human Baselines in Model Evaluations," arXiv, June 9, 2025, https://arxiv.org/abs/2506.13776.

130. Richard Ngo (@RichardMCNgo), "2. Provides signal across scales. Evals are often designed around a binary threshold (e.g. Turing Test). But this restricts the impact of the eval to a narrow time window around hitting it. Much better if we can measure (and extrapolate) orders-of-magnitude improvements." X, July 18, 2024, https://x.com/RichardMCNgo/status/1814049102197666209; Luca Righetti,

"Dangerous Capability Tests Should Be Harder," Planned Obsolescence, August 20, 2024, https://www.planned-obsolescence.org/dangerous-capability-tests-should-be-harder/. One avenue might be emphasizing so-called sophisticated capabilities in evaluations—see Ben Buchanan, *The Legend of Sophistication in Cyber Operations* (Belfer Center, January 2017), https://www.belfercenter.org/publication/legend-sophistication-cyber-operations.

131. Andrew Trask, et al., "Secure Enclaves for AI Evaluation," OpenMined, November 20, 2024, https://openmined.org/blog/secure-enclaves-for-ai-evaluation/; Lennart Heim, "A Trusted AI Compute Cluster for AI Verification and Evaluation," March 31, 2024, https://blog.heim.xyz/a-trusted-ai-compute-cluster/; Benjamin S. Bucknall and Robert F. Trager, *Structured Access for Third-Party Research on Frontier AI Models: Investigating Researchers' Model Access Requirements* (Centre for the Governance of AI, October 2023), https://cdn.governance.ai/Structured_Access_for_Third-Party_Research.pdf; and Bengio, et al., *International AI Safety Report*, sec. 3.4.3.

132. Hamin and Scott, *Hacking with AI*, 1.

133. Ashley Lin and Lennart Heim, "DeepSeek's Lesson: America Needs Smarter Export Controls," RAND, February 5, 2025, https://www.rand.org/pubs/commentary/2025/02/deepseeks-lesson-america-needs-smarter-export-controls.html; Samuel Hammond, "DeepSeek's Success Reinforces the Case for Export Controls," Foundation for American Innovation, January 30, 2025, https://www.thefai.org/posts/deepseek-s-success-reinforces-the-case-for-export-controls; *DeepSeek: A Deep Dive: Hearing Before the Subcommittee on Research and Technology of the Committee on Science, Space, and Technology*, 119th Cong. (2025) (statement of Gregory C. Allen, Director, Wadhwani AI Center, Center for Strategic and International Studies), https://republicans-science.house.gov/_cache/files/0/d/0d8e8a6c-2a09-413d-9d78-e21df078bc9e/EAB890C6D96E33647BC244A4EFCC742660C0EE09981A7B46E6BC1EC8C80F5386.gregory-c.-allen-testimony.pdf; and Dario Amodei, "On DeepSeek and Export Controls," January 2025, https://darioamodei.com/on-deepseek-and-export-controls.

134. *DeepSeek: A Deep Dive* (statement of Gregory C. Allen); Dylan Patel, Jeff Koch, and Sravan Kundojjala, "Fab Whack-A-Mole: Chinese Companies Are Evading U.S. Sanctions," SemiAnalysis, October 28, 2024, https://semianalysis.com/2024/10/28/fab-whack-a-mole-chinese-companies/; and Greg Allen, "MORE Export Controls: Foundry, DRAM, and Reflections on Biden," January 16, 2025, in *ChinaTalk*, produced by Jordan Schneider, podcast, 34:10, https://www.chinatalk.media/p/more-export-controls-foundry-dram.

135. Samuel Hammond and Erich Grunewald, "Spreadsheets vs. Smugglers: Modernizing the BIS for an Era of Tech Rivalry," Foundation for American Innovation, April 29, 2024, https://www.thefai.org/posts/spreadsheets-vs-smugglers-modernizing-the-bis-for-an-era-of-tech-rivalry; Barath Harithas, *Mapping the Chip Smuggling Pipeline and Improving Export Control Compliance* (Center for Strategic and International Studies, April 9, 2024), https://www.csis.org/analysis/mapping-chip-smuggling-pipeline-and-improving-export-control-compliance; Erich Grunewald and Tim Fist, "Comments on the Advanced Computing/Supercomputing IFR: Export Control Strategy & Enforcement for AI Chips," January 16, 2024, https://downloads.regulations.gov/BIS-2022-0025-0062/attachment_1.pdf; *DeepSeek: A Deep Dive* (statement of Gregory C. Allen); and *America's AI Action Plan*, 21.

136. *Bipartisan House Task Force Report on Artificial Intelligence*, sec. "Federal Preemption of State Law."

137. "Machine Learning Trends," Epoch AI.

138. Benton Martin and Jeremiah Newhall, "Technology and the Guilty Mind: When Do Technology Providers Become Criminal Accomplices," *Journal of Criminal Law and Criminology* 105, no. 1 (January 1, 2015), https://scholarlycommons.law.northwestern.edu/jclc/vol105/iss1/3; John Bandler and Antonia Merzon, *Cybercrime Investigations: A Comprehensive Resource for Everyone* (CRC Press, 2022); and Gregory Smith, et al., *Liability for Harms from AI Systems: The Application of U.S. Tort Law and Liability to Harms from Artificial Intelligence Systems* (RAND, November 20, 2024), https://www.rand.org/pubs/research_reports/RRA3243-4.html.

139. Open-weight models have several advantages. They lack external barriers to examining their inner workings, analyzing their behavior, developing new versions, or hosting them independently without the fear of losing access. A key distinction from closed-weight models is that open-weight models offer limited opportunities for their developers to monetize the societal value they generate. While holding developers of frontier models accountable for negative externalities can incentivize appropriate levels of harm mitigation, applying the same level of accountability to open-weight models risks unduly discouraging their development and release. With this said, open-weight models also have distinct risks—their release is effectively irreversible, for example. Also, safeguards can be fine-tuned away quickly and at relatively low cost. See *Dual-Use Foundation Models with Widely Available Model Weights* (National Telecommunications and Information Administration, July 30, 2024), https://www.ntia.gov/programs-and-initiatives/artificial-intelligence/open-model-weights-report.

140. Sayash Kapoor, et al., "On the Societal Impact of Open Foundation Models," arXiv, February 27, 2024, https://doi.org/10.48550/arXiv.2403.07918.

141. Holden Karnofsky, *A Sketch of Potential Tripwire Capabilities for AI* (Carnegie Endowment for International Peace, December 10, 2024), https://carnegieendowment.org/research/2024/12/a-sketch-of-potential-tripwire-capabilities-for-ai; Leonie Koessler, Jonas Schuett, and Markus Anderljung, *Risk Thresholds for Frontier AI* (Centre for the Governance of AI, June 20, 2024), https://www.governance.ai/research-paper/risk-thresholds-for-frontier-ai.

142. Sella Nevo, et al., *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models* (RAND, May 30, 2024), https://www.rand.org/pubs/research_reports/RRA2849-1.html.

143. Examples in other sectors include the Cybersecurity and Infrastructure Security Agency's (CISA's) CyberSentry Program for critical infrastructure entities (CISA, "CyberSentry Program," https://www.cisa.gov/resources-tools/programs/cybersentry-program) and the Department of Energy's Cybersecurity Risk Information Sharing Program (CRISP) for electricity industry entities (Department of Energy Office of Cybersecurity, Energy Security, and Emergency Response, "Cybersecurity Risk Information Sharing Program (CRISP)," https://www.energy.gov/sites/default/files/2021-12/CRISP%20Fact%20Sheet_508.pdf).

144. "OpenAI, Anthropic Sign Deals with US Govt for AI Research and Testing," Reuters, August 29, 2024, https://www.reuters.

com/technology/artificial-intelligence/openai-anthropic-sign-deals-with-us-govt-ai-research-testing-2024-08-29/.

145. Stages and descriptions reproduced from "Enterprise Tactics," MITRE ATT&CK, https://attack.mitre.org/tactics/enterprise/. Some selected examples draw on Mirsky, et al., "The Threat of Offensive AI to Organizations"; Andrey Anurin, et al., "Catastrophic Cyber Capabilities Benchmark (3CB): Robustly Evaluating LLM Agent Cyber Offense Capabilities," arXiv, November 2, 2024, https://arxiv.org/abs/2410.09114; and Buchanan, et al., *Automating Cyber Attacks*. These are commended as useful resources for readers seeking more detail on the applications of AI to cyber offense, along with: Hamin and Scott, *Hacking with AI*.

146. Adapted from Musser and Garriott, *Machine Learning and Cybersecurity*.

About the Center for a New American Security

The mission of the Center for a New American Security (CNAS) is to develop strong, pragmatic, and principled national security and defense policies. Building on the expertise and experience of its staff and advisors, CNAS engages policymakers, experts, and the public with innovative, fact-based research, ideas and analysis to shape and elevate the national security debate. A key part of our mission is to inform and prepare the national security leaders of today and tomorrow.

CNAS is located in Washington, D.C., and was established in February 2007 by cofounders Kurt M. Campbell and Michèle A. Flournoy. CNAS is a 501(c)3 tax-exempt nonprofit organization. Its research is independent and nonpartisan.

AMERICA'S EDGE 2025

The United States faces a rapidly changing global security landscape. Evolving technology, shifting alliances, and emerging threats require America to harness bold, innovative approaches. America's Edge is a Center-wide initiative featuring research, events, and multimedia for enhancing America's global edge.

CNAS Editorial

DIRECTOR OF STUDIES
Katherine L. Kuzminski

PUBLICATIONS & EDITORIAL DIRECTOR
Maura McCarthy

SENIOR EDITOR
Emma Swislow

ASSOCIATE EDITOR
Caroline Steel

CREATIVE DIRECTOR
Melody Cook

DESIGNER
Alina Spatz

Cover Art & Production Notes

COVER ILLUSTRATION
Mark Harris

PRINTER
CSI Printing & Graphics
Printed on an HP Indigo Digital Press

Center for a New American Security
1701 Pennsylvania Ave NW
Suite 700
Washington, DC 20006
CNAS.org
@CNASdc

Contact Us
202.457.9400
info@cnas.org

Center for a
New American
Security