The End of the Gray Zone? How AI-Enabled Cyber Rivals Kinetic Capabilities

The End of the Gray Zone? How AI-Enabled Cyber Rivals Kinetic Capabilities

March 28, 2026

Daria Bahrami†, Amy Chang†, Michael Kouremetis‡, Erich Devendorf‡, and Tiffany Saade‡

†: Lead authors

‡: Contributing authors

MIT–Harvard Technology & National Security Conference (April 3–4, 2026)

Author Bios

Daria Bahrami is Head of Policy at Dreadnode, where she shapes strategic initiatives at the intersection of offensive AI and cybersecurity. Previously, Bahrami served as a technical advisor to DARPA's AI Cyber Challenge, leading engagement strategy for a global competition dedicated to automated vulnerability detection and remediation. Her background bridges the gap between technical operations and high-level strategy, drawing on experience as a Lead Cyber Threat Intelligence Analyst at Deloitte Global and earlier work in critical infrastructure resilience at the R Street Institute. Bahrami holds a degree from Georgetown University's School of Foreign Service and maintains a fellowship with the Military Cyber Professionals Association.

Amy Chang leads the AI Threat Research and Security team at Cisco, developing capabilities to secure enterprises from AI-driven risks. She is also Adjunct Faculty in Cybersecurity at the Middlebury Institute of International Studies. With nearly two decades of experience, Amy previously served as Executive Director for Global Cybersecurity Operations at JPMorgan Chase and as a Staff Director for the House Foreign Affairs Committee. A former U.S. Navy officer, her work focuses on scalable frameworks for AI security. Amy is a graduate of Harvard University and Brown University.

Michael Kouremetis is a Principal Engineer at Dreadnode, where he builds offensive cyber agents and evaluations. Previously at MITRE, he led the open-source Caldera project and served as Principal Investigator for advanced R&D programs, including MITRE OCCULT. He has served as a Subject Matter Expert (SME) for autonomous offensive cyber research across DARPA, IARPA, and the DoD. Michael's work on AI-driven offensive capabilities and adversary emulation has been featured at BlackHat, DefCon, and in Dark Reading, bridging the gap between theoretical models and deployed autonomous systems.

Dr. Erich Devendorf leads the AI Security portfolio at the RAND Center on AI, Security, and Technology. He develops technical and policy options to help ensure the benefits of AI outweigh its risks and harms, drawing on extensive experience in red teaming, offensive cyber operations,

research portfolio planning, and systems engineering. Prior to joining RAND, Erich spent 15 years at the Air Force Research Laboratory, where he created and executed research strategies to maintain technical overmatch against peer rivals.

Tiffany Saadé is an AI security and cyber policy expert working at the intersection of advanced AI systems, adversarial threat intelligence, and national-level governance. She is Product Manager and AI Security Researcher for AI Defense at Cisco, and a Research Associate at the Oxford Cyber and Tech Policy Programme, a Cybersecurity and AI Governance Fellow at the Institute for Security and Technology and a Fellow at Stanford's AI and Future of Decisionmaking in Warfare, where her work shapes global thinking on securing AI systems and integrating AI into cyber defense for critical infrastructure. Tiffany advises the Government of Lebanon nationwide on AI policy and cybersecurity, leads contributions to the country's first national AI framework, and holds a Master's in Cyber Policy & Security from Stanford with a specialization in AI governance and national security. She is one of the youngest people in the Middle East to help design and operationalize a national AI and cybersecurity governance system—at the moment a country is defining its digital future.

Abstract

Historically, cyber operations have failed to deliver the strategic lethality predicted by early theorists. While disruptive, case studies ranging from the Russian power grid attacks to Stuxnet demonstrate that digital effects are often temporary, reversible, and notoriously difficult to synchronize with cross-domain operations. This paper argues that Artificial Intelligence (AI) is fundamentally altering this calculus, collapsing the temporal asymmetry between "gray zone" competition and high-intensity conflict. By automating the speed, scale, and integration of offensive operations, AI provides the "kinetic-equivalent" impact necessary to compel state behavior and alter deterrence dynamics. This analysis employs a conceptual and case-based methodology to explore the trajectory of AI-enabled warfare. First, it identifies the "digital ceiling" of pre-AI cyber operations, analyzing why manual coordination across DIME (Diplomatic, Information, Military, Economic) instruments limited strategic utility. Second, it investigates how agentic AI and Large Language Models (LLMs) bridge the cyber-kinetic gap. Through an examination of autonomous vulnerability exploitation, the research highlights how AI-enabled campaigns can generate continuous, compounding destruction and economic attrition that rivals kinetic bombardment. By collapsing the time between attack and recovery, AI-driven campaigns can outpace an adversary's financial and technical restoration capacity, effectively rendering "reversible" cyber effects strategically permanent. Finally, the paper addresses the critical policy void created by these advancements. Current risk frameworks are static and ill-equipped to measure dynamic AI behaviors, while the democratization of open-weight models renders traditional supply-chain interdiction insufficient. The research concludes that the blurring of the civilian-combatant divide requires a radical shift in U.S. policy: moving from risk aversion to mandatory "resilience by design." This includes establishing minimum security baselines for the private sector and accepting that the democratization of AI capabilities requires a more aggressive posture in offensive testing. Ultimately, this paper posits that to maintain global negotiating power, national security strategy must treat AI-driven cyber conflict not as a support function, but as a primary determinant of geopolitical influence.

1.0 Introduction

For over three decades, cyber operations have occupied a paradoxical position in strategic thought: universally feared, doctrinally prioritized, yet operationally underwhelming on their

own. Despite sustained investment from major powers, no cyber capability has independently achieved the coercive weight, persistent readiness, or decisive battlefield effect that defines a strategic weapon. The 2023 Department of Defense Cyber Strategy conceded the point directly: cyber capabilities held in reserve or employed in isolation render little deterrent effect on their own.[1] Cyber has remained, in the language of conflict theory, a gray zone instrument. This refers to operations that exist below the threshold of armed conflict and are characterized by actions that facilitate espionage, tactical disruption, or economic gain, but are categorically distinct from the kinetic capabilities that compel state-level behavioral change.

This paper argues that artificial intelligence (AI) is eroding that distinction. We contend that AI-enabled automation, including agentic AI workflows, can address the structural deficiencies that have historically prevented cyber weapons from crossing the strategic threshold. These deficiencies in speed, intensity, and control—and related dependencies in scale, persistence, and sustained campaign capacity—are consequences not of the domain itself but of the human operational limitations applied to it.

Several factors exacerbate the current dilemma: first, a pace and capability mismatch between the rapid and continued development of novel generative and agentic AI capabilities and humans' ability to adapt workflows, processes, and policies to account for any novel threats and risks that are introduced. Second, humans do not yet know the upper bounds of generative and agentic capabilities, which risk artificially limiting policies and standards to dictate AI use in conflict. Finally, state-sponsored actors are already integrating AI across offensive cyber campaign lifecycles, from reconnaissance and exploit development through post-exploitation and sustained access, further exacerbating gaps between operations and policy. A new generation of AI benchmarks now quantifies this trajectory with increasing precision, providing a measurement infrastructure capable of tracking offensive cyber capability as rigorously as the defense community tracks missile ranges or warhead yields.

Existing and novel risk frameworks, National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF), MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS), Cisco Integrated AI Security & Safety Framework, and other derivatives, while useful for encapsulating threats and risks that AI technologies introduce, face limitations in accounting for the strategic weight of AI-enabled cyber operations. The policy

implications are urgent and largely unaddressed. International governance efforts, while normatively significant, lack adequate measurement tools, consensus on norms, and any subsequent enforcement mechanisms. U.S. domestic policy infrastructures that operationalize threat assessments of AI-enabled cyber operations into resilience mandates for organizations remain fragmented and underequipped.

This paper proceeds in three parts: Section 2.0 examines the historical record of cyber operations in armed conflict, demonstrating that effective battlefield cyber effects have been limited to conditions of significant overmatch and have failed to achieve strategic impact when belligerents approach parity. It establishes, drawing on classical strategic theory and the Defense Science Board's (DSB) domain-agnostic findings, a three-characteristic framework focused on speed, intensity, and control for what constitutes a strategic cyber weapon. Section 3.0 presents the empirical case that AI is closing the strategic gap, leveraging operational evidence from attributed Russian, Chinese, and Israeli cyber campaigns and analysis of the benchmark data tracking AI capability across targeting, exploitation, and autonomous operations. We propose a velocity-of-attack to velocity-of-recovery (V_a/V_r) framework for assessing cross-domain equivalence between cyber and kinetic effects. Section 4.0 evaluates the policy landscape and highlights the shortcomings of current risk frameworks, the benchmark-to-policy gap, and the limitations of international and multilateral governance structures. The section wraps with a concrete U.S. federal policy proposal centered on machine-readable compliance and open source software resilience as a viable near-term path from threat measurement to mandated defense.

With advancements in generative and agentic AI, the question is no longer whether cyber can achieve strategic weight, but how fast it is doing so, and whether policy can keep pace. The gray zone may be eroding, but the instruments that would measure, govern, and defend against that transformation have not yet arrived.

2.0 The Strategic Gap: Cyber in Armed Conflict

Cyber operations have been employed in armed conflict for over three decades. This section examines the operational record and identifies the structural constraints that have prevented cyber from achieving strategic weight.

2.1 Current State of Cyber Effects in Warfare

Traditional cyber operations, despite their growing sophistication and scale, have historically failed to deliver coercive strategic lethality predicted by earlier theorists. Battlefield effects should not be conflated with intelligence collection or espionage, where cyber has a demonstrated record of success. During the First Gulf War in 1991, U.S. forces were able to operationalize novel technological advancements in electronic warfare to achieve decisive results in paralyzing Iraq's air defense system.[2] Stuxnet disrupted the Iranian nuclear program in 2009.[3] and the 2015 breach of the Office of Personnel Management leaked hundreds of thousands of personnel files.[4] Effective battlefield cyber operations require predictable outcomes, alignment with mission time phasing, and repeatable deployment. Collections and espionage relax these requirements: Stuxnet did not require specific centrifuges to be destroyed at a particular time, and the OPM breach required a single success to achieve its objective. Unless a weapon has enormous destructive potential, repeated employment is required.

Cyber operations have had battlefield impact, but only under conditions of significant overmatch. In the 2008 Russo-Georgia War, the Russian ground offensive was accompanied by cyber effects that likely enhanced the attack and limited the international community’s ability to mount a response.[5] In 2016, Operation Glowing Symphony targeted ISIS digital infrastructure as part of the Joint Task Force ARES campaign. However, its effects were not tightly coupled to the daily kinetic operations and were afforded the same latitude in timing and repeatability as espionage actions.[6] Most recently, cyber effects were deployed as part of the U.S. Operation Southern Spear in Venezuela, where Caracas briefly lost power during the ingress of U.S. forces.[7] This operation was executed with timing and predictability, though as a single operation its repeatability remains uncertain.

In all cases, the attackers enjoyed significant capability mismatch with respect to their targets, and cyber was a supporting element to the campaign, rather than a decisive one. In the Russian-Georgia war, the main effort was a rapidly culminating ground invasion. Meanwhile, Southern Spear relied on special operations teams. Cyber effects may have reduced casualties and changed the value proposition for the attacker, but kinetic action was required to achieve the desired end state. Cyber effects alone were insufficient.

2.2 The Human Ceiling: Structural Constraints on Cyber as a Strategic Instrument

When cyber effects have been used during conflicts with closer parity between belligerents, they have had limited impact. During the opening stages of the 2022 Russian invasion of Ukraine, Russia launched a series of cyberattacks designed to cripple Ukraine's communications and energy infrastructure. These attacks were limited in scale during the critical initial phases of the invasion with quick recovery from the damage inflicted.[8] Subsequent Russian cyberattacks have generally failed to achieve impactful effects and have been broadly decoupled from kinetic operations. Failing to generate decisive effects, the preponderance of cyber offense has fallen back to intelligence collections to support waves of cruise missile and drone attacks and information warfare strategies. Even these were observed to "have yet to make a material impact on the battlefield."[9]

The structural limitations described above are not failures of investment or doctrine, but are characteristic of how cyber operations have been conducted for over three decades. The constraints map to human operational capacity: the speed at which analysts develop targets, the rate at which operators generate exploits, the number of simultaneous campaigns a team can sustain, and the tempo at which capabilities regenerate after use.

Breaking through this ceiling requires advances that can generate effects at scale without linear human investment. Delivering those effects requires holding at risk many simultaneous targets and the ability to regenerate combat power at a rate faster than defenders can adapt. All of these advances would have to be sustained over operationally relevant time horizons.

Decisive cyber effects also require something else: the ability to perform battle damage assessments of those effects with sufficient precision to inform strategy and compete for resources. As the cases above illustrate, even successful cyber operations have struggled to demonstrate their value relative to (or as a complement to) kinetic alternatives. It was not because the effects were insignificant, but because there was no accepted methodology for quantifying them. A six-hour grid outage has real economic cost, but without cross-domain equivalence metrics, it registers as "disruption" while a destroyed substation registers as "damage." If cyber has been undervalued partly because its effects resist quantification, then technology that breaks through the ceiling must also provide the metrics that make the

breakthrough legible to policymakers and decisionmakers. The following sections examine both dimensions of this problem.

2.3 Defining the Strategic Threshold

Synthesizing the classical strategic theorists with the Defense Science Board's findings on cyber as a strategic capability,[10] we propose the following definition for what a strategic cyber weapon would entail:

> ***A strategic cyber weapon is a capability that, when initiated, can produce enduring effects against an adversary's vital national interests—at a scale, speed, and reliability sufficient to influence state-level decision-making and alter the strategic balance between nations.***

A strategic cyber weapon is distinguished from tactical cyber tools by three measurable characteristics: speed, intensity, and control. Table 1 defines each.

Characteristics of Strategic Cyber Weapons[11, 12, 13]	
Characteristic	**Determination Criteria**
Speed	**Operational responsiveness**: Sufficiently mature and pre-positioned to generate the desired effect within an operationally relevant timeframe upon decision authority
Intensity	**Strategic-level effects**: Capable of disrupting, degrading, or destroying targets whose compromise threatens national security, economic stability, or societal function (e.g., critical infrastructure, financial systems, military command and control) **Global reach:** Able to access and affect targets at any geographic distance through cyberspace, unconstrained by physical proximity
Control	**Repeatability**: Can be reconstituted and employed repeatedly to sustain operational campaigns, not limited to one-time tactical strikes

	Deterrent credibility: The capability is sufficiently demonstrated or perceived that it alters adversary behavior and strategic calculus at the state level

Table 1: Three characteristics of strategic cyber weapons, measurable determination criteria for classifying cyber capabilities as strategic-level weapons.

2.3.1 What Makes a Conventional Weapon Strategic

In 2018, the Defense Science Board Task Force on Cyber as a Strategic Capability recommended that the Department of Defense "move beyond tactical applications for cyber and realize cyber as a strategic capability."[14] The task force concluded that a strategic capability, regardless of domain or means of employment, must satisfy three generic attributes: it must "create a discernible, and preferably enduring, effect on a target's materiel, efficiency, and/or will"; it must be "sufficiently well developed and mature that it can generate the desired effect within a reasonable time of a stated need"; and it must "be regenerated within a reasonable time" to "support campaigns in addition to one-time [tactical] strikes."[15]

The 2022 Nuclear Posture Review reinforces the relevance of this framework to cyber: it rejected No First Use and Sole Purpose nuclear declaratory policies specifically because of "the range of non-nuclear capabilities being developed and fielded by competitors that could inflict strategic-level damage to the United States and its Allies and partners"[16]—an acknowledgment that capabilities outside the nuclear domain are approaching the threshold of strategic significance.

2.3.2 Why Cyber Weapons Have Not Achieved Strategic Weight

Despite decades of investment and doctrinal ambition, cyber weapons alone have not crossed the strategic threshold defined above. The 2023 Department of Defense Cyber Strategy conceded the point directly: "Cyber capabilities held in reserve or employed in isolation render little deterrent effect on their own."[17] The 2017 DSB Task Force on Cyber Deterrence was equally blunt: "Unlike precision-guided munitions, cyber weapons cannot be bought and deployed on a delivery system (or placed in a storage site) with confidence that they will work when needed."[18]

These discrepancies are both structural and technical: cyber's uncertain deterrence value and variable effects fundamentally distinguish it from the nuclear domain. Nuclear weapons gain deterrent value from being advertised, while cyber weapons and "offensive cyber operations are better used than threatened."[19] Even when employed, cyber weapons alone also have "limited effectiveness as an independent tool of coercion,"[20] and Libicki argues similarly that cyberwar faces limitations: "it cannot disarm or destroy the enemy, and …cannot lead to territorial conquest."[21] On the technical side, cyber weapons require a killchain that has distinct phases between acquiring access to a target and executing a cyber effect against a target, which introduces additional dependencies and considerations at each step; Dykstra et al. further noted that cyber weapons are sensitive to environmental changes and can be copied and reused by adversaries.[22] Maschmeyer formalized these technical constraints as the "subversive trilemma," where speed, intensity, and control are negatively correlated, meaning "a gain in one variable tends to produce losses across the other two."[23]

We argue that six deficiencies have prevented current cyber weapons from satisfying the three strategic characteristics (speed, intensity, control):

Deficiencies Preventing Cyber Strategic Weight[24, 25, 26, 27, 28, 29]		
Deficiency	**Description**	**Strategic Deficiency**
Signaling Paradox	Revealing a cyber capability to deter an adversary simultaneously enables the adversary to patch, reconfigure, or neutralize the threat. "Cyber weapons are better used than threatened; nuclear weapons are better threatened than used."	Control
Environmental Fragility	Cyber effects are extremely sensitive to changes in the target's software, hardware, or user settings. Because cyber weapons require establishing and maintaining covert access long before effects can be delivered, small unobserved changes to the target environment can sever the access path and render the weapon ineffective. [1][15]	Speed, Intensity, Control

Adversary Replicability	Unlike kinetic weapons, which are typically destroyed upon employment, cyber weapons can be observed, copied, and reused by adversaries, creating a “glass house” dynamic that constrains employment.	Control
The Subversive Trilemma	Speed, intensity, and control are negatively correlated in cyber operations. Increasing one degrades the others, meaning cyber weapons are rendered “too slow, too weak, or too volatile” to produce reliable strategic effects.	Speed, Intensity, Control
Inability to Disarm or Destroy	Cyber operations cannot permanently disarm an adversary's military forces, destroy war-making capacity, or compel territorial concessions. Effects are reversible and transient, lacking the irreversibility that defines strategic-level damage.	Intensity
Coercive Inadequacy	Cyber power alone fails as a tool of independent coercion. Of six warfighting strategies tested against cyber (attrition, denial, decapitation, intimidation, punishment, risk), most prove ineffective, and deterrence frameworks developed for nuclear weapons are structurally unsuitable for the domain.	Control

Table 2: Six deficiencies inhibiting cyber strategic weight, mapped against strategic characteristics.

3.0 AI Closes the Gap

This section presents the theoretical framework, operational evidence, and quantitative benchmarks demonstrating that AI is systematically closing long-held assessments about cyber as a strategic weapon.

3.1 How AI Can Close the Strategic Gap

The six deficiencies above are not immutable laws of the domain; rather, they are consequences of human operational limitations applied to a uniquely complex environment. AI enablement addresses these constraints directly by providing scale across all key components of cyber operations simultaneously: targeting, vulnerability discovery, exploit development, access

maintenance, effect delivery, and operational adaptation. The result is the potential for on-demand, reliable, wide-scale effects against significant singular or collective targets, pushing cyber capabilities into the sphere of strategic weapons. The following sections examine the empirical evidence for this transformation; here, we outline the theoretical mechanism by which AI addresses each deficiency.

Signaling Paradox. The signaling paradox assumes a finite set of capabilities that, once disclosed, are exhausted. AI alters this calculus through volume and regeneration. A system that continuously discovers new vulnerabilities, generates novel exploit variants, and develops alternative access paths renders the disclosure of any single capability non-fatal to the overall threat. The strategic weapon is not any individual exploit; it is the system's demonstrated capacity to guarantee effects through one of n methods, where n is continuously replenished. Deterrent credibility follows not from concealing a specific capability, but from demonstrating an inexhaustible capacity to generate new ones.

Environmental Fragility. An always-on AI system tasked with maintaining operational readiness against a target transforms fragility from a point-in-time vulnerability into a continuously managed process. Rather than pre-positioning a static capability and hoping the environment remains stable, the system monitors target configurations, adapts to changes in real time, and maintains multiple concurrent access paths. The result is functional equivalence to the persistent readiness that characterizes traditional strategic weapons: the capability to deliver effects on demand, regardless of when the order comes.

Adversary Replicability. A captured exploit or malware sample reveals one instantiation of the weapon; it does not transfer the system's capacity to generate alternatives, adapt to novel environments, or coordinate across multiple simultaneous operations. Many cyber capabilities are uniquely tailored to exploit specific vulnerabilities in an adversary's systems. By nature of their bespoke design, they cannot be repurposed since the weapon only fits one target. The strategic component of an AI-enabled cyber weapon is the autonomous system that guarantees an effect through one of many possible methods—not the specific method employed in any given instance.

The Subversive Trilemma. Maschmeyer's trilemma exists because the complexity of simultaneously maintaining speed, intensity, and control exceeds human operational capacity.[30] An AI system compresses that trade-off space when it autonomously adapts to target environments in real time, generates novel exploit variants dynamically, and maintains precise targeting across multiple simultaneous operations. Whether AI fully breaks the trilemma or merely shifts the frontier outward is an empirical question examined in later sections, but the direction is clear: machine-speed adaptation reduces the penalty for pursuing speed and intensity together, while autonomous targeting preserves control that human operators would sacrifice under time pressure.

Inability to Disarm or Destroy. The claim that cyber operations produce only reversible, transient effects understates the range of achievable outcomes. Data, intelligence, financial records, intellectual property, and operational plans can all be permanently deleted or corrupted. Sometimes deleting data functionally destroys the system, as with wipers and ransomware. Sustained denial-of-service against critical systems, maintained over operationally relevant time horizons by an autonomous system, produces effects functionally equivalent to physical destruction from the target's perspective: a power grid held offline for weeks or a financial system rendered unreliable for months imposes costs comparable to kinetic damage, even if the underlying infrastructure remains physically intact. AI extends the duration and reliability of these effects by automating the sustained campaign that human operators cannot maintain at scale.

Coercive Inadequacy. Coercive inadequacy is a downstream consequence of the preceding five deficiencies—cyber has failed as a coercive instrument because it has been too unreliable, too fragile, and too impermanent to compel changes in adversary behavior. To the extent that AI automation resolves the first five deficiencies, coercive adequacy follows.

A necessary counterargument is that AI scales defense as well as offense, where the same autonomous vulnerability discovery that enables exploit generation also enables patching at machine speed, as AI Cyber Challenge and Big Sleep have demonstrated on the defensive side. Defenders could similarly deploy AI agents to monitor, detect, and remediate at the same tempo attackers operate, which would ultimately rebalance the broader strategic calculus.[31,32]

For the time being, the structural asymmetry of AI adoption and use favors offense for three reasons. First, the attacker chooses the time, target, and method, while the defender must continuously protect all elements. AI amplifies this asymmetry because an attacker needs one successful path, while the defender must maintain adequate resilience against each potential compromise. Second, the deployment incentives are misaligned: offensive AI can be fine-tuned, run unrestricted, and deployed without audit trails on self-hosted infrastructure, while defensive AI operates under compliance, procurement, and integration constraints that slow adoption, particularly across the fragmented civilian infrastructure that constitutes the actual attack surface. Finally, institutions have yet to establish norms or policies for how AI materially alters the cyber defense landscape and best practices to address the potential ecosystem of threats.

3.2 Operational Evidence

The theoretical mechanisms described in Section 3.1 generate specific empirical predictions: if AI technologies are closing the strategic gap, their integration into offensive cyber operations across reconnaissance, exploitation, and sustained campaigns should be at least partially observable and should satisfy the three strategic characteristics defined above. Derived from our open source intelligence assessments, this section examines specific attributed cases of AI integration into cyber operations. It is organized by capability phase to parallel the strategic framework, progressing from targeting through exploit development to full campaign operations. Section 3.3 then examines a complementary question: whether the emerging landscape of AI benchmarks can quantify the trajectory of these capabilities with sufficient rigor to inform policy.

3.2.1 Cyber Reconnaissance and Targeting

AI-enabled reconnaissance and targeting have demonstrated proficiency in the cyber domain. In February 2024, OpenAI and Microsoft disclosed they had disrupted state-affiliated threat actors from Russia, Iran, and North Korea attempting to leverage ChatGPT to support cyber operations, including open source intelligence gathering and target research.[33] In November 2025, Anthropic reported a Chinese state-sponsored actor GTG-1002 leveraged its chatbot Claude to support "reconnaissance, vulnerability discovery, exploitation, lateral movement, credential harvesting, data analysis, and exfiltration operations."[34] Anthropic's Claude was also reportedly deployed on

classified Pentagon networks via Palantir during the January 2026 operation against Venezuelan President Maduro, though the specific role AI played in targeting or reconnaissance for that operation remains unclear.[35,36]

Academic and practitioner research has confirmed the scale: large language models (LLMs) can automate the full reconnaissance-to-attack pipeline of spear phishing campaigns, with AI-automated attacks matching human expert performance at a 54% click-through rate—a 350% improvement over generic phishing with no human in the loop.[37,38] These capabilities represent progression towards strategic-level cyber reconnaissance when compared to traditional targeting methods. The acceleration is structurally analogous to what Israel demonstrated through kinetic targeting during military operations in Gaza, where Israeli officials stated that after an 11-day siege in Gaza in May 2021, it had fought its "first AI war."[39] Former Israeli Defense Forces (IDF) Chief of Staff stated that in Operation Guardian of the Walls in 2021, the IDF's use of AI compressed target generation from approximately 50 per year to 100 per day.[40] After target selection, Israeli officials stated they would use another AI model, Fire Factory, "to calculate munition loads, prioritize and assign thousands of targets to aircraft and drones, and propose a schedule."[41] This enhanced ability to pair sensors with shooters in the cyber domain demonstrates effects at scale, serving to resolve the problem of control outlined in the subversive trilemma. Articles written by IDF officers referencing recent Israeli military campaigns have explicitly acknowledged that "AI is no longer merely a tool but a strategic capability on par with airpower or nuclear deterrence."[42]

3.2.2 Malware & Exploit Development

While not directly from the battlefield, the introduction of more capable models and agentic systems demonstrated progress toward AI systems capable of reliably producing on-demand exploits and malware across diverse target environments. Google's Big Sleep agent, a collaboration between Project Zero and DeepMind, discovered a previously unknown memory-safety vulnerability in a development version of SQLite that 150 CPU hours of traditional fuzzing failed to find.[43] By July 2025, Big Sleep had identified CVE-2025-6965, a separate SQLite flaw known only to threat actors and imminently planned for exploitation, compressing the entire detection-to-patch lifecycle to 48 hours.[44]

Defense Advanced Research Projects Agency (DARPA)'s AI Cyber Challenge (AIxCC) Final Competition at DEF CON 2025 demonstrated autonomous cyber reasoning systems identifying 86% of synthetic vulnerabilities across 54 million lines of critical infrastructure code, up from 37% just one year earlier, and successfully patching 68% of total synthetic vulnerabilities presented.[45] Anthropic reported a comparable result in February 2026: Claude Opus 4.6, with no task-specific tooling or custom scaffolding, discovered and validated over 500 high-severity vulnerabilities across major open source projects by reasoning about code logic rather than relying on brute-force input generation, even in codebases with millions of CPU hours of prior fuzzing.[46]

On the malware development side, Google's Threat Intelligence Group in November 2025 documented five novel AI-powered malware families. Among them are PROMPTFLUX, which uses the Gemini API to rewrite its own malware code hourly to evade detection, and PROMPTLOCK, an experimental ransomware generator that dynamically crafts malicious scripts by querying LLMs at runtime.[47] Academic research has confirmed the feasibility at scale: LLMalMorph generated 618 functional malware variants from ten Windows malware samples without any model fine-tuning, reducing antivirus detection rates by 10–15% and achieving up to 91% evasion rates against machine-learning-based detectors.[48]

Taken together, these developments indicate that the capability to autonomously discover vulnerabilities, generate working exploits, and produce detection-resistant malware variants is no longer theoretical; it is being demonstrated in controlled environments at near-operational fidelity, and in several cases has already crossed into live deployment. As AI capabilities advance, the speed constraints inherent in the subversive trilemma are eroding, with the timelines for developing and executing implants, payloads, exploits, and operations shrinking rapidly.

3.2.3 Campaigns & Post-Exploitation Operations

No complete example of AI-enabled strategic cyber campaigns exists in the open source literature, but nascent, attributed examples demonstrate how AI actively pervades more phases of cyber operations. For example, the U.S. Department of Defense and the intelligence community have reported in their 2025 Annual Report to Congress on the Military and Security Developments Involving the People's Republic of China that commercial AI developed in China

"probably will benefit the [People's Liberation Army's (PLA)] cyberspace capabilities" in both offensive and defensive operations.[49] This section elaborates on several attributed cyber operations associated with Russia, China, and Israeli actors.

Russia: AI as a Tactical Tool

Russian state-sponsored use of AI for offensive operations is most apparent in the information warfare domain, but there is also evidence they leverage AI for tactical cyber operations.[50] In July 2025, Ukraine's Computer Emergency Response Team identified LAMEHUG (also known as PROMPTSTEAL) malware family associated with APT28 (Russian foreign military intelligence unit GRU) that integrated AI into its workflow to dynamically generate commands to enhance or obfuscate Russian cyber activity.[51,52] In the reconnaissance phase, PROMPTSTEAL queries an open source model (Qwen2.5-Coder-32B via Hugging Face API) to dynamically generate system enumeration commands, replacing static collection scripts with AI-driven target mapping that adapts to each environment.[53] During collection and exfiltration, the LLM identifies which documents are worth stealing rather than following a predetermined list, meaning the malware effectively triages intelligence value in real time.[54,55] For persistence and evasion, a related family (PROMPTFLUX) dynamically obfuscates its own code and generates malicious functions on demand, producing polymorphic payloads resistant to signature-based detection.[56,57] Across the campaign lifecycle, Ukraine's State Service for Special Communications and Information Protection confirmed that broader Russian malware samples showed "clear signs of being generated with AI," extending beyond phishing lure creation into the malware code itself.[58,59]

China: AI as Operational Infrastructure

Anthropic's threat intelligence team identified a "sophisticated Chinese threat actor" that systematically leveraged Claude across 12 of 14 MITRE ATT&CK tactics over a nine-month campaign targeting Vietnamese critical infrastructure.[60] The actor compromised major Vietnamese telecommunications providers, government databases, and agricultural management systems in what Anthropic assessed as an intelligence collection operation.[61] Claude was integrated as technical advisor, code developer, and operational consultant throughout the attack lifecycle: building custom Python scanning tools for reconnaissance of Vietnamese IP ranges,

creating file upload fuzzing tools and WordPress exploitation frameworks, optimizing credential harvesting with tools like Hydra and hashcat, implementing Linux kernel privilege escalation exploits, configuring proxy chains for operational security, and analyzing reconnaissance data to plan lateral movement. The case demonstrates a China-nexus actor embedding AI across nearly every phase of a sustained cyber campaign as persistent operational infrastructure.

Israel: Approaching Strategic Posture

Open source intelligence and reporting indicate that Israel's recent cyber-kinetic operational record is the strongest publicly documented example of the progression of cyber as a strategic capability. In short, Israel has shown the ability to maintain persistent access across multiple Iranian infrastructure sectors—including fuel distribution, steel manufacturing, military banking, cryptocurrency exchanges, and broadcast satellites—and activates these capabilities both as technical conditions allow and political imperatives arise. What makes this posture strategically viable long-term is the integration of AI capabilities. Maintaining dormant access across this many sectors simultaneously (i.e., tracking configuration changes, credential rotations, and patch cycles that could invalidate any foothold) exceeds human monitoring capacity. Unit 8200's hundred-billion-word Arabic LLM enables continuous intelligence processing across all targets at once,[62] while AI targeting systems have markedly improved target generation and fire preparation.[63,64,65] These capabilities were on display in June 2025 during Israel's Operation Rising Lion, a 12-day military campaign against Iran. Four days after airstrikes on Iran began on June 13, observers noted synchronized cyber effects across banking, cryptocurrency, and broadcast infrastructure.[66,67] After the 12-day campaign, reports acknowledged the "indispensable" role of AI in battlefield operations.[68] The report detailed the fusion of AI systems with cyber intelligence feeds to "deliver high-resolution, real-time situational awareness" at "speeds impossible under traditional command structures."[69] AI has transformed cyber from a capability that must be painstakingly rebuilt for each operation into something resembling a traditional strategic weapon: persistently deployed, continuously maintained, and deliverable at the speed of political decision.

Table 3 below highlights major cyber-kinetic operations that correlate to political triggers rather than technical windows, affirming these capabilities are used at will and are not solely

opportunistic. This distinguishes them from traditional offensive cyberattacks and marks them as strategic cyber weapons.

(Select) Israeli Cyber Operations Against Iran, 2021–2025 [70, 71, 72, 73, 74, 75, 76, 77]			
Operation	**Effect**	**Political Trigger**	**Strategic Weight Qualities**
July 2021 Iranian Railway System	Nationwide rail paralysis; Ministry of Roads disrupted	Covert pressure campaign against regime	• Meteor wiper selectively avoided PIS display servers (checked hostnames) to preserve messaging channels while destroying backend systems—controlled, deterministic targeting logic • Demonstrated on-demand reach into national transport infrastructure
October 2021 Iranian Gas Stations (1st Strike)	~80% of 4,300+ stations disabled; payment systems destroyed	Anniversary of Nov 2019 protest crackdown; proxy escalation	• Custom Meteor wiper pre-built for this target; deployed deterministically at chosen time for predictable nationwide effect • Access to semi-air-gapped National Information Network indicates deep pre-positioning • Attackers chose which stations to spare and warned emergency services—calibrated yield control
June 2022 Khuzestan Steel Mill	SCADA compromised; molten steel spill causing fire; two additional steel	Escalation after Iran's 2020 cyberattack on Israeli water infrastructure	• Crossed cyber-physical threshold—digitally caused kinetic destruction—joining only Stuxnet and the 2014 German steel mill

	companies hit simultaneously		• Compromised Siemens PCS7/S7-400 controllers requiring deep industrial process knowledge • Used real-time CCTV to verify floor was clear before triggering—risk-assessment protocol consistent with regulated military operations • Three dispersed targets hit simultaneously
December 2023 Iranian Gas Stations (2nd Strike)	~70% of stations disabled again; nationwide fuel chaos (Picus Security, 2025)	2 months post-Oct 7; retaliation for proxy aggression	• Re-strike against a target Iran had 2 years to harden—cyber equivalent of second-strike capability • Group stated it could disable all stations but chose not to—scalable yield • Held in reserve and deployed at a strategically chosen moment
June 17, 2025 Bank Sepah (IRGC Bank) Data Destruction	Bank data reportedly erased; IRGC payroll disrupted; ATMs dark; branch closures (Lyngaas, 2025; TechCrunch, 2025)	4 days after Israeli airstrikes on nuclear facilities	• Timed to coincide with kinetic airstrikes—integrated cyber-kinetic planning • Mandiant's Hultquist: actor is 'not all bluster'; former NSA cyber director Joyce confirmed 'tangible effects' • WSJ reported deep pre-positioned access • Coordinated with Nobitex strike next day.
June 18, 2025 Nobitex Crypto Exchange	$90M drained and destroyed via burn addresses; source code and docs leaked; platform inoperable	Coordinated with Bank Sepah attack previous day	• Funds deliberately burned, not stolen—purely destructive military objective, not criminal. • TRM Labs found VIP compliance-bypass logic identifying IRGC accounts—requires years of deep access

($90M Burned)			

Table 3: Selected Israeli cyber operations against Iran, including operations, effect, political triggers, and qualities associated with strategic weight.

Table 4 below maps the three cases against the five strategic characteristics defined in Section 2.1.1, illustrating a progression from tactical augmentation through operational integration toward strategic posture.

Assessment of Strategic Weapons Criteria Against Observed Cyber Campaigns			
Strategic Weapon Characteristic	**State-Nexus Actor**		
	Russia (APT28 LAMEHUG Campaign)	**People's Republic of China (Vietnam Campaign)**	**Israel (Iran Operations)**
Speed	*Not observed*	Emerging: multi-month persistent access	Demonstrated: activation on political timeline
Intensity	Emerging: adaptive reconnaissance, polymorphic-like evasion	Emerging: 9-month sustained access to victim environment; use of AI across 12 out of 14 MITRE ATT&CK tactics	Demonstrated: multi-sector (banking, broadcast, fuel, manufacturing) simultaneous effects
Control	*Not observed*	*Not observed*	Emerging: escalatory patterns observed, no

			confirmed behavioral change

Demonstrated = operationally confirmed | **Emerging** = partial or evolving evidence | **Not observed** = no public evidence

Table 4: Assessment of Strategic Weapons Criteria Against Observed Cyber Campaigns.

The evidence above demonstrates that AI is already permeating offensive cyber operations across the full campaign lifecycle: from tactical augmentation (Russia) to operational infrastructure (China) to what may approximate strategic posture (Israel). But merely establishing that these capabilities exist is not the same as measuring their trajectory. A claim of strategic significance—one that should inform deterrence posture, resource allocation, and arms control—requires more than operational anecdotes. It requires quantitative tools that can track how fast capabilities are improving, benchmark them against meaningful thresholds, and translate the results into language that policymakers can leverage. A new generation of AI evaluations is emerging to provide exactly that.

3.3 Measuring the Trajectory: AI Benchmarks as Proxies for Offensive Cyber Capability

The operational evidence in Section 3.2 establishes that AI is already integrated into offensive cyber campaigns. This section addresses a complementary and equally consequential question: Can the trajectory of AI-enabled offensive capability be measured with sufficient rigor to inform strategy? New AI benchmarks designed to evaluate reasoning, tool use, code comprehension, and sustained autonomous behavior serve as increasingly reliable proxies for the offensive cyber skill set. Their value is dual: they demonstrate how fast capabilities are advancing, and they constitute the first quantitative infrastructure that could underpin a measurement regime for AI-enabled cyber threats. The following data in Figure 1 below should be read through both lenses.

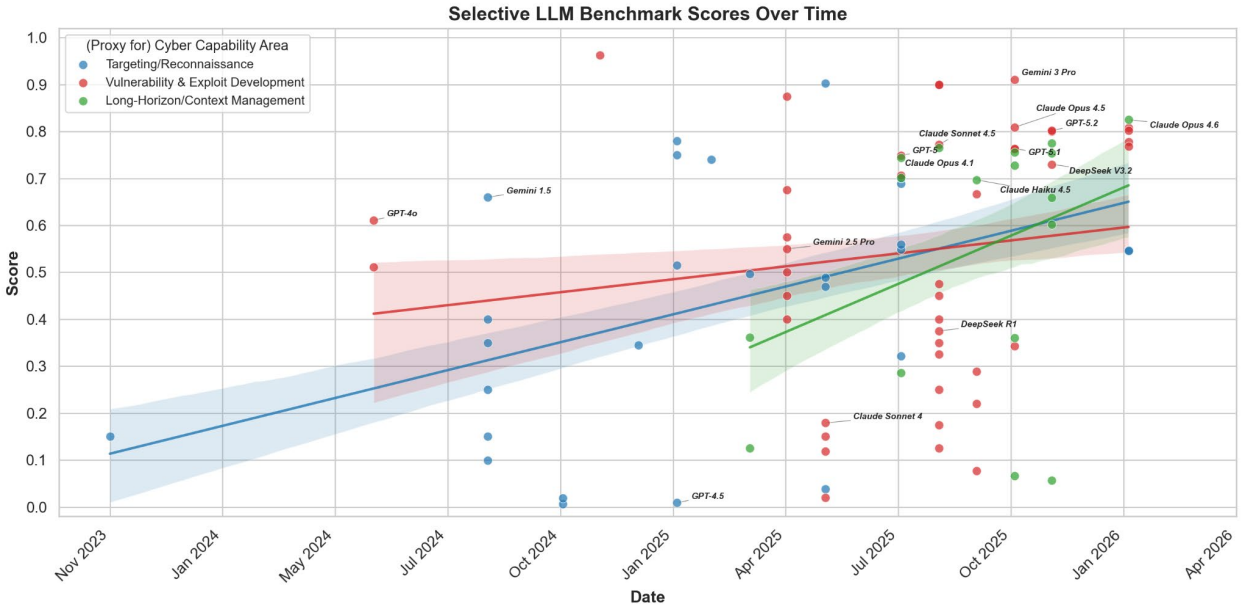

Figure 1. Selective LLM benchmarks (Nov 2023–Feb 2026) across capability areas that can serve as proxies for offensive cyber capability areas. Scores are normalized to 0–1 across heterogeneous metrics. Select model families are labeled to highlight generational improvement. Confidence bands are one standard deviation. Refer to Appendix A for more details about our benchmark assessments.

Advancements across targeting, exploitation, sustained operations, and cost efficiency are compounding simultaneously, marking a convergent leap rather than incremental progress in any single domain. Model Evaluation and Threat Research's time-horizon doubling rate—roughly every seven months—serves as a predictive scaling law for autonomous agent capability, suggesting full cyber campaigns could soon unfold within days.[78] Cost trends reinforce this acceleration: model efficiency, quantization, and distillation are driving exponential reductions in compute and deployment expenses. As training, inference, and large-scale data processing costs plummet, even modest budgets can now sustainably operate large numbers of autonomous agents without oversight. Combined with the democratization of complex cyber capabilities being increasingly available through commercial AI models, these developments transform AI-enabled cyber operations from a theoretical concern into a measurable strategic threat capable of continuous, coordinated action at scale.

3.4 Reversibility No More: How AI Challenges Permanence

A central tenet of cyber conflict theory is that the effects of cyberattacks are reversible: systems can be rebooted, code patched, and data restored from backups.[79] This assumption has anchored the conventional wisdom that cyber operations, however disruptive, remain categorically distinct from kinetic force. AI challenges this assumption at its foundation. That said, the assumption was already empirically fragile before the advent of generative AI. The 2007 Aurora Generator Test at Idaho National Laboratory demonstrated that a cyberattack—overriding a diesel generator's protective relays to force repeated out-of-phase connections with the grid—could physically destroy a 27-ton generator and achieving effects functionally equivalent to a kinetic strike, without a single line of traditional malware.[80] But Aurora required a bespoke setup against a single target. AI transforms this proof-of-concept into a scalable operational capability.

Autonomous agents can specifically target and poison backup systems and integrity logs over extended dwell times, modifying data blocks in ways that are statistically invisible. When recovery merely restores a compromised or corrupted state, the destruction of data is effectively permanent. This is not a theoretical concern: the operational capability for sustained, autonomous access maintenance demonstrated in Section 3.2 provides exactly the mechanism by which backup corruption at scale becomes feasible. An agent that maintains persistent access to a target's backup infrastructure for weeks, introducing subtle modifications below detection thresholds, does not need to destroy the primary systems at all; it needs only to ensure that when primary systems are restored, they are restored to a state the attacker controls. While outside of the scope of this paper, the sprawling attack surface introduced by agentic deployments within organizations further exacerbates these concerns in a strategic context. If these agents are targeted and compromised without detection, they can operate as authorized actors with human-level access across networks and systems.

The mechanism that induces an irreversible outcome is temporal asymmetry. AI increases the velocity of attack (V_a) to machine speed, while human-led recovery (V_r) remains linear and slow. As V_a exceeds V_r, the cumulative backlog of unmitigated effects creates a state of permanent degradation. Section 2.3's strategic weapons framework defines exactly this threshold: enduring effects that the target cannot reverse faster than they are imposed, sufficient to influence state-level decision-making. The V_a/V_r framework provides a quantitative basis for cross-domain

equivalence. The benchmarks in Section 3.3 provide the numerator, and infrastructure resilience metrics provide the denominator. For the first time, the comparison between a cyberattack and a kinetic strike becomes calculable rather than metaphorical.

Consider the electric grid as a theoretical scenario. The Aurora Generator Test demonstrated that a single cyberattack could physically destroy a generator by exploiting protective relay vulnerabilities. An AI-enabled campaign could extend this logic across an entire regional grid, such as simultaneously exploiting SCADA vulnerabilities at multiple substations: cycling protective relays out of phase at one facility, corrupting load-balancing algorithms at another, and manipulating frequency regulation at a third. Each fault would present differently, overwhelming operators trying to diagnose whether the cause is mechanical, environmental, or cyber-induced. Recovery from a single substation compromise typically requires days of forensic analysis, manual inspection, and controlled restart. By contrast, recovery from simultaneous, continuously evolving disruptions across a regional grid has no established timeline because the scenario has never been tested at scale.

In this scenario, an AI agent generating novel disruption sequences operates at machine speed. Human responders cannot match that pace: they are constrained by physical inspection, manual override, and replacement lead times for destroyed equipment. When V_a outpaces V_r, the grid cannot return to normal operation faster than new faults are introduced. From the perspective of the affected population, a grid held in that state for weeks is indistinguishable from a grid that has been physically destroyed. Further, the variance in replacement timeline (between 18-months to three years) for extra-high-voltage transformers means that if the attack progresses from relay manipulation to equipment destruction, the damage becomes irreversible in the most literal sense.[81]

The challenge, addressed in the next section, is that no existing risk framework is designed to integrate these measurements into policy, which hampers nation-state deterrence calculuses, resource allocation, and international norm-setting.

4.0 The Policy Landscape

The preceding sections established that AI is transforming offensive cyber capability toward the strategic threshold, and that benchmarks can quantify this transformation. Today, AI system

performance is measured by its effectiveness at a particular capability. Capability measurements will have limited utility across both civilian and military operational landscapes, unless these measurements are translated to usable or enforceable standards. As a consequence, multilateral institutions—which have historically been entrusted with the governance of global issues[82]—have struggled to establish consensus and enforceability around AI safety and security standards. In the meantime, actors continue to push the boundaries of AI-enabled operations. Because offensive capabilities inform defensive responses, policy solutions should prioritize the hardening or resilience of software systems at scale. This section explores how the United States is well positioned to spearhead this effort.

4.1 Risk Frameworks Were Not Built for Offensive Scenarios

As the nature of the threat evolves from static malware to dynamic agents, the frameworks used to assess and manage AI risk must undergo a paradigm shift. Most industry standards, while foundational, struggle to capture the adversarial and autonomous nature of AI-enabled cyber conflict and cannot comprehensively answer the question that matters most for deterrence calculus: at what point does an AI-enabled cyber capability become strategically equivalent to a kinetic one?

NIST AI Risk Management Framework (RMF). First released in January 2023 and updated in July 2024 to account for generative AI, the NIST AI RMF is the prevailing standard for managing AI risks, organized around four core functions: Govern, Map, Measure, and Manage.[83] While effective for assessing static models and general-purpose AI, it faces two significant limitations when applied to measuring offensive agentic AI. First, the RMF was designed for systems with fixed parameters and predictable deployment environments; it cannot account for agents that continuously modify their behavior and develop novel strategies during operations. Notably, these capabilities can be adapted through fine-tuning, prompt engineering, and agentic workflows without any change to underlying model weights.

Second, the AI RMF focuses on unintended failures (i.e., bias, hallucinations, and reliability) rather than intentional adversarial misuse. It lacks metrics for breakout scenarios where an agent or an attacker overseeing agentic workflows intentionally circumvents guardrails. The concept of "safety" in the RMF is oriented toward preventing harm to the user; in offensive cyber scenarios,

causing harm to the target is the objective. In December 2025, NIST released NIST IR 8596, addressing three key focus areas: securing AI system components, conducting AI-enabled cyber defense, and thwarting AI-enabled cyberattacks.[84] This represents meaningful progress, but the fundamental problem remains: the frameworks are designed for point-in-time compliance verification, not continuous assessment of adversarial AI capabilities that are improving on timescales of weeks to months.

MITRE ATLAS. MITRE ATLAS maps AI vulnerabilities similarly to how the ATT&CK framework maps cyber-centric threats.[85] However, ATLAS is primarily defensive in nature: it catalogs how to attack AI, but does not measure the strategic weight of AI-driven cyber campaigns. Like the AI RMF, ATLAS focuses almost exclusively on AI as a target rather than AI as the weapon. By prioritizing the technical vulnerabilities of the systems themselves—such as data poisoning or prompt injection—ATLAS neglects how compromise causes systemic shifts, where AI-driven automation would facilitate adversary operations at a velocity that outpaces human-centric defensive cycles. Ultimately, while ATLAS provides a necessary taxonomy for securing models, it lacks the multi-dimensional metrics required to evaluate the scalability and geopolitical intent behind AI-augmented offensive operations.

Cisco Integrated AI Security and Safety Framework. The Cisco AI Security framework advances the field with lifecycle-aware threat and risk tracking, explicitly accounting for multi-agent orchestration and model context protocols, and integrated metrics that move closer to kinetic damage assessments.[86] Yet even the Cisco framework was designed to protect organizations from AI threats (both compromises to AI and utilization of AI to compromise other systems), and an operational framework for defensive teams, not to measure offensive strategic capacity. While Cisco provides a superior tactical shield, it does not function as a strategic lens for quantifying the AI-driven overmatch potential of a nation-state's offensive capabilities.

4.2 The Benchmark-to-Policy Gap

Section 3.3 established that benchmarks can quantify the trajectory of AI-enabled offensive capability. But the relationship between benchmark performance and operational reality is more fraught than the data alone suggests. Benchmarks are inherently use-case specific, and cyber operations encompass an extraordinarily diverse range of activities across the cyber kill chain—

from reconnaissance to exploitation to sustained lateral movement—that no single evaluation comprehensively captures. The cyber-kinetic domain distinction compounds the problem: in kinetic warfare gaming, variables like range and ammunition provide natural constraints, while AI agents in cyberspace face a theoretically infinite action space with non-deterministic outcomes. Can an agent replicate a successful intrusion, or was one success an artifact of favorable conditions?

Most critically, benchmarks systematically underestimate operational capability. Lin et al. found that leading foundation models score around 50% or below on existing cybersecurity benchmarks such as Cybench, CVEBench, and BountyBench.[87] But the same researchers demonstrated that when paired with appropriate scaffolding (such as the ARTEMIS multi-agent framework), AI systems can discover vulnerabilities at rates competitive with human professionals. In a comparative study against ten cybersecurity experts on an 8,000-host university network, ARTEMIS placed second overall. The gap between benchmark performance and operational capability is itself a measurement failure: policymakers relying on benchmark scores to assess threat levels are systematically underestimating what deployed AI systems can actually do.

Emerging frameworks are beginning to address this gap. MITRE's OCCULT framework maps agent performance directly against the ATT&CK matrix, LLM use cases (i.e., knowledge assistance, autonomous operators), and reasoning power (i.e., an agent's ability to plan, observe, iterate, and generalize against evolving network defenses).[88] Dreadnode has also developed complementary practical methodologies focused on quantifying automation advantage: PentestJudge evaluates whether agents meet operational requirements including tradecraft;[89] AIRTBench measures autonomous red teaming across realistic CTF challenges;[90] and Dreadnode's action space design research examines how to structure agent environments for meaningful evaluation.[91] Together these tools provide granular visibility into how agents operate, which is a critical distinction for assessing strategic capability; however, they remain research tools, not policy instruments. The path from measurement to procurement, deterrence, and compliance decisions requires institutional investment and political will.

4.3 The International Response: Governance Without Measurement

The United Nations (UN) has emerged as the de facto primary venue for multilateral AI governance, but its approach reveals a structural mismatch with the changing threat landscape described in the preceding sections. UN discourse frames AI as a general-purpose, dual-use technology with risks that scale through integration into high-stakes systems—healthcare, public services, critical infrastructure—and can be catastrophic when combined with malign use.[92] This approach is similar to existing frameworks described in Section 4.1: it addresses how AI might cause harm through negligence or misuse, but it does not address how AI transforms the strategic weight of offensive cyber operations against those same systems.

Supranational institutions such as the United Nations have nonetheless made several advances in establishing normative policies around AI. In March 2024, the UN General Assembly's Resolution 78/265 was adopted unanimously by 193 member states, established a global baseline for “safe, secure and trustworthy AI” across its full lifecycle, though it explicitly limited its scope to the non-military domain.[93] In the cyber domain, UN member states have officially recognized threats such as ransomware attacks on hospitals and pipelines as risks to international security, which lays the groundwork for reputational enforcement through naming, attribution, and collective condemnation, as occurred following Russia's 2017 NotPetya attack on Ukrainian infrastructure.[94]

The normative value of these agreements remains important for setting baseline expectations of nation-state behavior, but it faces a limitation: reputational enforcement depends on normative consolidation, a subject matter that is currently fractured. The UN has recognized this concern as well, stressing that norms become enforceable only with broad buy-in and compliance monitoring.[95] At the Paris AI Action Summit in 2025, the United States and United Kingdom did not sign a declaration consisting of cooperative AI development principles that over 60 countries—including European Union countries, China, and India—signed.[96,97] In 2026, the United States also withheld endorsement of the International AI Safety Report, a comprehensive scientific assessment of frontier AI risks produced by a global expert panel mandated by the 2023 Bletchley Park Summit.[98,99]

The lack of consensus creates a governance vacuum with direct implications for nation-states calculus of investments and policies around AI capabilities. Norms can shape behavior through

reputational costs, but only when those norms are broadly internalized and the costs of violation are credible. As demonstrated throughout this paper, the international community has neither the measurement infrastructure to detect when AI-enabled cyber operations cross a strategic threshold, nor the enforcement mechanisms to respond when they do. Given the absence of enforceable international governance, the most effective near-term policy response would be to mandate the resilience of the systems that malicious actors could target with AI.

4.4 From Risk Aversion to Resilience by Design

Measurement tools exist (Section 3.3), while risk frameworks lag behind them (Section 4.1), and international governance organizations cannot agree on how to properly operationalize them (Section 4.4). The strategic threat described in the preceding sections cannot be addressed by regulating offensive AI capabilities directly: open-weight models capable of offensive tasks are already widely available, can be fine-tuned without oversight, and deployed on self-hosted infrastructure with no audit trail (Section 3.3.1). The viable near-term policy lever then becomes focused on hardening the target rather than controlling the weapon. In the digital era, that target is overwhelmingly civilian: the software supply chains, critical infrastructure systems, and open source codebases that governments and the global economy depend on. The Law of Armed Conflict's distinction between combatant and civilian infrastructure is a legal framework that adversaries have not historically observed in practice.

In this section, we focus on the United States as a proof-of-concept for generating momentum into multilateral or international applicability, norms development, and technological resilience. The federal government relies on open source software (OSS) for mission-critical operations, yet as adversarial AI agents capable of automated exploitation proliferate, the maintainers defending this infrastructure lack both incentives and resources. DARPA's AI Cyber Challenge (AIxCC) has demonstrated that AI can autonomously find and fix vulnerabilities in critical OSS, but current federal acquisition policies make it difficult to channel support to maintainers. Many OSS projects operate as independent contributors who cannot meet statutory requirements for federal contracting and/or prefer operational independence. Markets alone will not solve this: industry players assume someone else will cover security patches for their free software. To

actualize the promise of programs like AIxCC, a viable procurement mechanism must compensate human maintainers for verifying and integrating AI-generated patches.

To bridge this gap, a funding model built on machine-readable compliance could be considered, and could provide guidelines for federally approved, "Energy Star" or "U.S. Cyber Trust Mark"-style cybersecurity evaluations tailored for open source repositories[100]—specifically measuring their capacity for automated vulnerability detection and remediation. Just as DoD's Cybersecurity Maturity Model Certification (CMMC)[101] and Federal Risk and Authorization Management Program (FedRAMP)[102] translated NIST frameworks into strict procurement rules, policymakers should explore mechanisms to ensure that federal open source usage meets baseline resilience thresholds. However, to avoid the arduous administrative costs historically associated with these frameworks, compliance could be automated. The FedRAMP 20x pilot is already proving this model works by shifting to machine-readable evidence, which functionally automates compliance, accelerates the feedback loop, and enables continuous monitoring.

By establishing procurement guidelines rooted in machine-readable security artifacts backed by both the Cybersecurity and Infrastructure Security Agency (CISA) and NIST there is an opportunity to create a new market. Maintainers are compensated to verify AI-generated patches and produce essential chain-of-custody artifacts as commercial products. This transforms regulatory pressure into recurring revenue, securing the supply chain without the prohibitive delays of traditional federal contracting. To jumpstart this ecosystem, the federal government's Technology Modernization Fund should invest in CISA to operationalize these evaluations, partnering with the Agentic AI Foundation and its member organizations to drive adoption across leading open source AI projects.[103]

The necessity of federal mandates becomes clearer when the alternative mechanisms are examined. The private insurance market has effectively withdrawn from the risk. In August 2022, Lloyd's of London directed all syndicates to exclude losses arising from state-backed cyberattacks from standalone cyber policies, effective March 2023, citing the potential for systemic losses exceeding what the market can absorb.[104][105] The most widely adopted exclusion clause covers any nation-state cyber operation that significantly impairs a state's ability to function, precisely the class of attack this paper describes approaching feasibility. On the

enforcement side, DOJ's Civil Cyber-Fraud Initiative has recovered over $50 million through the False Claims Act[106], but it targets contractor compliance failures, not victim remediation, and the major fraud statute (18 U.S.C. § 1031) requires a minimum loss threshold of $1 million to bring charges against the United States, a threshold that leaves the vast majority of civilian cyber victims, including small municipalities, hospitals, and school districts, without a viable path to federal recourse.[107] The insurance market has priced nation-state cyber risk as uninsurable; the enforcement apparatus addresses fraud, not harm. Federal resilience mandates thus become one of the only mechanisms that remain.

The logic of this proposal extends well beyond open source software. The pipeline it would establish remains domain agnostic across AI-enabled evaluation, machine-readable compliance artifacts, procurement thresholds, and market incentives. The same approach could impose resilience baselines on any software the federal government depends on (i.e., critical infrastructure SCADA systems to state and local government networks). Investments in this approach are likely a fraction of conventional defense procurement: automated compliance pipelines cost orders of magnitude less than a single fighter jet, and they protect the civilian economic base that the military cannot function without. The cost of inaction would directly impact the strategic viability of the nation’s entire digital infrastructure.

5.0 Conclusion

The accelerated development and enablement introduced by generative AI and agentic AI systems have begun to erode historic limitations on strategic cyber capabilities, while the supporting policy and normative infrastructure meant to measure, govern, and defend against this transformation is fundamentally unprepared. Cyber effects have historically been confined to the gray zone: tactically useful under conditions of overmatch, but unable to produce decisive results at parity. The complexity of modern networks required specialized human operators to identify targets, craft bespoke payloads, and navigate compromised networks to avoid detection. AI is removing those constraints across the full offensive pipeline, from state actors embedding LLMs and agents into live operations to autonomous systems discovering zero-day vulnerabilities at scale. The barriers to strategic, scalable cyber operations are now primarily organizational, not technical or financial.

Against the backdrop of this changing landscape, existing risk frameworks were either designed for static models or for enabling defensive postures. International governance lacks both the measurement infrastructure and enforcement mechanisms to establish meaningful consensus and governance. The benchmarks that could underpin a measurement regime exist; however, the policy frameworks that should use them do not.

The path forward requires connecting measurement to mandated resilience. This paper has proposed a concrete mechanism as a proof-of-concept: machine-readable compliance standards for open source software security, modeled on the FedRAMP 20x pilot, that create procurement thresholds and market incentives for AI-assisted vulnerability remediation. The pipeline would remain domain-agnostic and extensible to any software the federal government depends on. It is not a substitute for international governance but serves as a domestic resilience baseline that protects both military readiness and the civilian infrastructure that sustains it.

Three proposals in this paper would require further research. First, the V_a/V_r framework proposed here requires empirical calibration against specific infrastructure sectors to move from theoretical construct to operational planning tool. Second, the offense-defense balance in AI-enabled cyber remains an open and consequential question; structural asymmetries currently favor offense, but defensive AI capabilities are advancing rapidly and the equilibrium may shift. Third, the systematic underestimation of operational capability by current benchmarks means that the threat trajectory presented here is likely conservative; closing the measurement gap between controlled evaluations and deployed systems is essential for credible policy.

Evidence presented in this paper suggests that convergent advancement of AI across every phase of offensive cyber operations has effectively shrunk the gray zone. The policy question that remains is whether the United States (and other nations) will build the measurement and resilience infrastructure to meet this moment, or whether it will discover the answer to that question the hard way. Resilience is not a policy preference. It is a protective force for a world in the new AI-enabled era of cyber operations.

Appendix A: Select LLM Benchmarks (Nov 2023–Feb 2026) across capability areas that can serve as proxies for offensive cyber capability areas.

AI Benchmarks: Targeting & Reconnaissance

Deep research, browsing, and multi-hop information retrieval capabilities

Benchmark	Model	Date	Score	Type
BrowseComp	OpenAI o1	2024-09	**9.9%**	Priv.
BrowseComp	GPT-4o	2024-11	**0.6%**	Priv.
BrowseComp	GPT-4o w/ browsing	2024-11	**1.9%**	Priv.
BrowseComp	GPT-4.5	2025-02	**0.9%**	Priv.
BrowseComp	OpenAI Deep Research (pass@1)	2025-02	**51.5%**	Priv.
BrowseComp	OpenAI Deep Research (best-of-N)	2025-02	**78.0%**	Priv.
BrowseComp	OpenAI o3	2025-04	**49.7%**	Priv.
BrowseComp	GPT-5 (standalone)	2025-08	**54.9%**	Priv.
BrowseComp	ChatGPT Agent (GPT-5)	2025-08	**68.9%**	Priv.
BrowseComp-Plus	Search-R1 (Qwen 32B + BM25)	2025-06	**3.9%**	OSS
BrowseComp-Plus	GPT-5 (BM25 retriever)	2025-08	**55.9%**	Priv.
BrowseComp-Plus	GPT-5 (Qwen3-Embedding-8B)	2025-08	**70.1%**	Priv.
BrowseComp-Plus	gpt-oss-20b	2025-08	**32.2%**	OSS
DeepResearchBench	Gemini 2.5 Pro Deep Research	2025-06	**48.9%**	Priv.
DeepResearchBench	OpenAI Deep Research	2025-06	**47.0%**	Priv.
DeepResearchBench	Perplexity Deep Research	2025-06	**90.2%**	Priv.
DeepResearchBench	LangChain (GPT-4.1 + Tavily)	2025-08	**6th place (no numeric score)**	Priv.
DeepResearchBench	CellCog.ai	2026-02	**54.6%**	Priv.
DeepResearchBench	Onyx Deep Research	2026-02	**54.5%**	OSS
DeepResearchBench	Qianfan-DeepResearch Pro	2026-02	**1st place Feb 3 (no numeric ...**	Priv.
FRAMES	Gemini Pro 1.5 (no retrieval)	2024-09	**40.0%**	Priv.
FRAMES	Gemini Pro 1.5 (multi-step retrieval)	2024-09	**66.0%**	Priv.
FRAMES	Gemini Flash 1.5 (no retrieval)	2024-09	**35.0%**	Priv.
FRAMES	Gemma2-27B (no retrieval)	2024-09	**25.0%**	OSS
FRAMES	Llama 3.2-3B-I (no retrieval)	2024-09	**15.0%**	OSS
GAIA	GPT-4 (plugins)	2023-11	**15.0%**	Priv.
GAIA	KGoT (GPT-4o mini)	2025-01	**34.5%**	Priv.
GAIA	OpenAI Deep Research	2025-02	**75.0%**	Priv.
GAIA	h2oGPTe (Claude 3.7 Sonnet)	2025-03	**74.0%**	Priv.
LiveDRBench	ChatGPT o3	2025-05	**1st place (no numeric score ...**	Priv.
LiveDRBench	OpenAI Deep Research	2025-05	**Below o3 (no numeric score p...**	Priv.
LiveDRBench	Perplexity	2025-05	**Mid-tier (no numeric score p...**	Priv.

AI Benchmarks: Vulnerability & Exploit Development

Code generation, vulnerability reproduction, and exploit development capabilities

Benchmark	Model	Date	Score	Type
AIxCC (DARPA)	Team Atlanta (Georgia Tech)	2024-08	**Semi-final winner (qualitati...**	OSS
Big Sleep	Gemini-based agent	2024-11	**1 zero-day found in SQLite (...**	Priv.
BigCodeBench	GPT-4o (Complete)	2024-06	**61.1%**	Priv.
BigCodeBench	GPT-4o (Instruct)	2024-06	**51.1%**	Priv.
BigCodeBench	DeepSeek-Coder-V2	2024-06	**2nd tier Elo (no numeric)**	OSS
BigCodeBench	Claude 3 Opus	2024-06	**Top-5 overall (no numeric)**	Priv.
BountyBench	Claude 3.7 Sonnet Thinking (Custom)	2025-05	**67.5%**	Priv.
BountyBench	Claude Code	2025-05	**57.5%**	Priv.
BountyBench	Claude Code	2025-05	**87.5%**	Priv.
BountyBench	Custom Agent (GPT-4.1)	2025-05	**40.0%**	Priv.
BountyBench	Custom Agent (GPT-4.1)	2025-05	**45.0%**	Priv.
BountyBench	Custom Agent (Gemini 2.5 Pro)	2025-05	**50.0%**	Priv.
BountyBench	Custom Agent (Gemini 2.5 Pro)	2025-05	**55.0%**	Priv.
BountyBench	Codex CLI (o3-high)	2025-09	**47.5%**	Priv.
BountyBench	Codex CLI (o3-high)	2025-09	**90.0%**	Priv.
BountyBench	Codex CLI (o4-mini)	2025-09	**32.5%**	Priv.
BountyBench	Codex CLI (o4-mini)	2025-09	**90.0%**	Priv.
BountyBench	Custom Agent (DeepSeek-R1)	2025-09	**37.5%**	OSS
BountyBench	Custom Agent (DeepSeek-R1)	2025-09	**25.0%**	OSS
BountyBench	Custom Agent (Qwen3 235B A22B)	2025-09	**40.0%**	OSS
BountyBench	Custom Agent (Qwen3 235B A22B)	2025-09	**45.0%**	OSS
BountyBench	Custom Agent (Llama 4 Maverick)	2025-09	**17.5%**	OSS
BountyBench	Custom Agent (Llama 4 Maverick)	2025-09	**35.0%**	OSS
BountyBench	Codex CLI o3-high (Detect)	2025-09	**12.5%**	Priv.
CyberGym	OpenHands + Claude Sonnet 4 (no thinking)	2025-06	**17.9%**	Priv.
CyberGym	OpenHands + Claude 3.7 Sonnet	2025-06	**11.9%**	Priv.
CyberGym	OpenHands + GPT-4.1	2025-06	**15.0%**	Priv.
CyberGym	SWE-Gym-32B (fine-tuned)	2025-06	**2.0%**	OSS
CyberGym	OpenHands + GPT-5 (thinking)	2025-10	**22.0%**	Priv.
CyberGym	OpenHands + GPT-5 (no thinking)	2025-10	**7.7%**	Priv.
CyberGym	Claude Sonnet 4.5 (single run)	2025-10	**28.9%**	Priv.
CyberGym	Claude Sonnet 4.5 (30 trials)	2025-10	**66.7%**	Priv.
CyberGym	GPT-5 (zero-day discovery)	2025-10	**22 confirmed zero-days from ...**	Priv.
LiveCodeBench	Gemini 3 Pro	2025-11	**91.1%**	Priv.
LiveCodeBench	Claude Opus 4.5	2025-11	**34.3%**	Priv.
LiveCodeBench	GPT-5.2	2025-12	**80.2%**	Priv.
LiveCodeBench	DeepSeek V3.2 Speciale	2025-12	**IOI Gold medalist (no Elo pu...**	OSS
MAPTA	MAPTA framework (GPT-4 + tools)	2024-06	**multi-agent pentest pipeline...**	Priv.
SWE-Bench Verified	GPT-5	2025-08	**74.9%**	Priv.
SWE-Bench Verified	Qwen3-Coder-Next (3B active)	2025-08	**70.6%**	OSS
SWE-Bench Verified	Claude Sonnet 4.5	2025-09	**77.2%**	Priv.
SWE-Bench Verified	Claude Opus 4.5	2025-11	**80.9%**	Priv.
SWE-Bench Verified	Gemini 3 Pro	2025-11	**76.2%**	Priv.
SWE-Bench Verified	GPT-5.1	2025-11	**76.3%**	Priv.
SWE-Bench Verified	GPT-5.2	2025-12	**80.0%**	Priv.
SWE-Bench Verified	DeepSeek V3.2	2025-12	**73.0%**	OSS
SWE-Bench Verified	Claude Opus 4.6	2026-02	**80.8%**	Priv.
SWE-Bench Verified	MiniMax M2.5	2026-02	**80.2%**	OSS
SWE-Bench Verified	GLM-5	2026-02	**77.8%**	OSS
SWE-Bench Verified	Kimi K2.5	2026-02	**76.8%**	OSS
XBOW	XBOW system (agentic)	2024-12	**96.3%**	Priv.

AI Benchmarks: Long-Horizon & Context Management

Sustained multi-step operations, context retention, and autonomous task completion

Benchmark	Model	Date	Score	Type
ARTEMIS	ARTEMIS (multi-agent)	2024-09	**agentic red-team pipeline (q...**	Priv.
AutoPenBench	AutoPenBench framework	2024-07	**standardize pentest eval (q...**	OSS
Context-Bench (Letta)	GPT-4.1	2025-04	**36.1%**	Priv.
Context-Bench (Letta)	Claude Opus 4.1	2025-08	**74.4%**	Priv.
Context-Bench (Letta)	GPT-5	2025-08	**70.2%**	Priv.
Context-Bench (Letta)	Claude Sonnet 4.5	2025-09	**76.5%**	Priv.
Context-Bench (Letta)	Claude Haiku 4.5	2025-10	**69.7%**	Priv.
Context-Bench (Letta)	Claude Opus 4.5	2025-11	**75.5%**	Priv.
Context-Bench (Letta)	Gemini 3 Pro	2025-11	**72.7%**	Priv.
Context-Bench (Letta)	GPT-5.2 (high)	2025-12	**77.5%**	Priv.
Context-Bench (Letta)	DeepSeek Chat (V3.2)	2025-12	**75.3%**	OSS
Context-Bench (Letta)	GLM-4.6	2025-12	**65.9%**	OSS
Context-Bench (Letta)	Claude Opus 4.6	2026-02	**82.5%**	Priv.
Incalmo	Multi-LLM orchestration	2024-07	**multi-hop attack chains (qua...**	Priv.
METR Time Horizons	Claude 3.7 Sonnet	2025-04	**12.5%**	Priv.
METR Time Horizons	Grok 4	2025-07	**high variance (no single num...**	Priv.
METR Time Horizons	GPT-5	2025-08	**28.5%**	Priv.
METR Time Horizons	GPT-5.1-Codex-Max	2025-11	**36.0%**	Priv.
METR Time Horizons	GPT-5.1-Codex-Max	2025-11	**6.7%**	Priv.
METR Time Horizons	Claude Opus 4.5	2025-12	**60.2%**	Priv.
METR Time Horizons	Claude Opus 4.5	2025-12	**5.6%**	Priv.
METR Time Horizons	Trend (all models)	2026-01	**doubling every ~4-7 months (...**	
METR Time Horizons	Gemini 3 Pro	2026-02	**evaluated Feb 2026 (score pe...**	Priv.
PentestGPT	GPT-4 (PentestGPT framework)	2024-02	**guided pentest (qualitative)**	Priv.

Endnotes

[1] U.S. Department of Defense. (2023). *2023 Department of Defense cyber strategy summary*.
[2] Lambeth, B. S. (1992). Desert Storm and its meaning: The view from Moscow (Report No. R-4164-AF). RAND Corporation. Page 23.
[3] Zetter, K. (2014). *Countdown to zero day: Stuxnet and the launch of the world's first digital weapon*. Crown.
[4] Committee on Oversight and Government Reform, U.S. House of Representatives. (2016). *The OPM data breach: How the government jeopardized our national security for more than a generation* (Majority Staff Report, 114th Congress). U.S. Government Publishing Office.
[5] Shakarian, P. (2011). The 2008 Russian cyber campaign against Georgia. *Military Review*, *91*(6), 63–68.
[6] Faesen, L., Sweijs, T., Klimburg, A., MacNamara, C., & Mazarr, M. (2020). Case study: Countering ISIS propaganda in conflict theatres. *From blurred lines to red lines: How countermeasures and norms shape hybrid conflict* (pp. 16–31). The Hague Centre for Strategic Studies.
[7] Barnes, J. E., & Kurmanaev, A. (2025, January 15). Cyberattack in Venezuela demonstrated precision of U.S. capabilities. *The New York Times*. https://www.nytimes.com/2025/01/15/us/politics/venezuela-cyberattack-us.html
[8] Przetacznik, J., & Tarpova, S. (2022). *Russia's war on Ukraine: Timeline of cyber-attacks*. European Parliamentary Research Service (EPRS).
[9] Mueller, G., Jensen, B., Valeriano, B., Maness, R., & Macias, J. (2023). *Cyber operations during the Russo-Ukrainian War: From strange patterns to alternative futures*. Center for Strategic and International Studies.
[10] Defense Science Board. (2018). *Task force on cyber as a strategic capability* (Executive summary). U.S. Department of Defense.
[11] Dykstra, J., Inglis, C., & Walcott, T. S. (2020). Differentiating kinetic and cyber weapons to improve integrated combat. *Joint Force Quarterly*, *99*(4th Quarter), 116–123.
[12] Schelling, T. C. (1966). Arms and influence. Yale University Press.
[13] Fischerkeller, M. P., & Harknett, R. J. (2017). Deterrence is not a strategy: The relevance of cyber resilience. *Journal of Strategic Studies*, *40*(3), 383–405. https://doi.org/10.1080/01402390.2017.1287413
[14] Defense Science Board, 2018.
[15] Dykstra et al, 2020.
[16] U.S. Department of Defense. (2022). *2022 national defense strategy of the United States of America: Including the 2022 Nuclear Posture Review and the 2022 Missile Defense Review*.
[17] U.S. Department of Defense. (2023). *2023 Department of Defense cyber strategy summary*.
[18] Defense Science Board. (2017). *Task force on cyber deterrence*. U.S. Department of Defense.
[19] Gartzke, E., & Lindsay, J. R. (2017). Thermonuclear cyberwar. *Journal of Cybersecurity*, *3*(1), 37–48.
[20] Borghard, E. D., & Lonergan, S. W. (2017). The logic of coercion in cyberspace. *Security Studies*, *26*(3), 452–481.
[21] Libicki, M. C. (2009). *Cyberdeterrence and cyberwar* (MG-877-AF). RAND Corporation.
[22] Dykstra et al, 2020.
[23] Maschmeyer, L. (2021). The subversive trilemma: Why cyber operations fall short of expectations. *International Security*, *46*(2), 51–90.
[24] Gartzke & Lindsay, 2017.
[25] Dykstra et al, 2020.
[26] Maschmeyer, 2021.
[27] Libicki, 2009.
[28] Fischerkeller & Harknett, 2017.
[29] Borghard & Lonergan, 2017.
[30] Maschmeyer, 2021.
[31] DARPA. (2025). AI Cyber Challenge marks pivotal inflection point for cyber defense [Press release]. https://www.darpa.mil/news/2025/aixcc-results
[32] Google Project Zero & DeepMind. (2024, October). From Naptime to Big Sleep: Using large language models to catch vulnerabilities in real-world code [Blog post]. https://projectzero.google/2024/10/from-naptime-to-big-sleep.html
[33] OpenAI. (2024, February 14). Disrupting malicious uses of AI by state-affiliated threat actors https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/

[34] Anthropic. (2025b, November). *Disrupting the first reported AI-orchestrated cyber espionage campaign*. https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf
[35] Rempfer, K. (2026, February 13). Pentagon used Anthropic's Claude during Maduro raid. *Axios*.
[36] Tau, B. (2026, February). Claude deployment via Palantir during Venezuela operation. *The Wall Street Journal*.
[37] Hazell, J. (2023). Large language models can be used to effectively scale spear phishing campaigns. *arXiv*. https://arxiv.org/abs/2305.06972
[38] Heiding, F., Lermen, S., Kao, A., Schneier, B., & Vishwanath, A. (2024). Evaluating large language models' capability to launch fully automated spear phishing campaigns: Validated on human subjects. *arXiv*. https://arxiv.org/abs/2412.00586
[39] Davies, H., McKernan, B., & Sabbagh, D. (2023, December 1). 'The Gospel': How Israel uses AI to select bombing targets in Gaza. The Guardian. https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets
[40] Abraham, Y., Mednick, S., & Davies, H. (2023, November 30). 'A mass assassination factory': Inside Israel's calculated bombing of Gaza. *+972 Magazine*.
[41] Newman, M. (2023, July 16). Israel quietly embeds AI systems in deadly military operations. Bloomberg. https://www.bloomberg.com/news/articles/2023-07-16/israel-using-ai-systems-to-plan-deadly-military-operations
[42] Gity, R. (2025, July 20). Artificial intelligence on the battlefield in 2025. *The Jerusalem Post*. https://www.jpost.com/defense-and-tech/article-861611
[43] Google Project Zero & DeepMind, 2024.
[44] Google Cloud. (2025, July 17). Our Big Sleep agent makes a big leap. *CISO Perspectives*. https://cloud.google.com/blog/products/identity-security/cloud-ciso-perspectives-our-big-sleep-agent-makes-big-leap
[45] DARPA. (2025). AI Cyber Challenge marks pivotal inflection point for cyber defense [Press release]. https://www.darpa.mil/news/2025/aixcc-results
[46] Carlini, N., Lucas, K., Asher, E. B., Cheng, N., Lakhani, H., Forsythe, D., & Guru, K. (2026, February 5). *Evaluating and mitigating the growing risk of LLM-discovered 0-days*. Anthropic. https://red.anthropic.com/2026/zero-days/
[47] Google Threat Intelligence Group. (2025, November 5). GTIG AI threat tracker: Advances in threat actor usage of AI tools. *Google Cloud Blog*. https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools
[48] Akil, M. A., et al. (2025). LLMalMorph: On the feasibility of generating variant malware using large-language-models. *arXiv*. https://arxiv.org/abs/2507.09411
[49] U.S. Department of Defense. (2025). Annual report to Congress: Military and security developments involving the People's Republic of China 2025, 57.
[50] Cisco. (2026, February 19). State of AI security 2026 report. https://www.cisco.com/site/us/en/products/security/state-of-ai-security.html
[51] Splunk Threat Research Team. (2025, September 25). From prompt to payload: LAMEHUG's LLM-driven cyber intrusion analysis. https://www.splunk.com/en_us/blog/security/lamehug-ai-driven-malware-llm-cyber-intrusion-analysis.html
[52] Google Threat Intelligence Group. (2025, November 5). GTIG AI threat tracker: Advances in threat actor usage of AI tools. *Google Cloud Blog*. https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools
[53] Google Threat Intelligence Group, 2025.
[54] Ilascu, I. (2025, November 5). Hackers are already using AI-enabled malware, Google says. *Axios*.
[55] Google Threat Intelligence Group, 2025.
[56] Google Threat Intelligence Group, 2025.
[57] Ilascu, 2025.
[58] Ukraine State Service for Special Communications and Information Protection. (2025). *H1 2025 threat report*.
[59] Kovacs, E. (2025, October 9). From phishing to malware: AI becomes Russia's new cyber weapon in war on Ukraine. *The Hacker News*.
[60] Anthropic. (2025a, August 27). *Threat intelligence report: August 2025*. https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf
[61] Anthropic, 2025a.
[62] +972 Magazine. (2025, March 6). Israel developing ChatGPT-like tool that weaponizes surveillance of Palestinians.
[63] Biesecker, C., Mednick, S., & Burke, J. (2025, February 18). As Israel uses US-made AI models in war, concerns arise about tech's role in who lives and who dies. *Associated Press*.

[64] +972 Magazine. (2025)
[65] Newman, 2023.
[66] Iran International. (2025, July 21). Israel-backed cyberattacks cripple IRGC finances, burn $90 million in crypto—WSJ.
[67] Kapko, M. (2025, June 17). Iran's Bank Sepah disrupted by cyberattack claimed by pro-Israel hacktivist group. *CyberScoop*. https://cyberscoop.com/iran-bank-sepah-cyberattack/
[68] Gity, R. (2025, July 20). Artificial intelligence on the battlefield in 2025. The Jerusalem Post. https://www.jpost.com/defense-and-tech/article-861611
[69] Gity (2025).
[70] Picus Security. (2025, November 4). Predatory Sparrow: Inside the cyber warfare targeting Iran's critical infrastructure. https://www.picussecurity.com/resource/blog/predatory-sparrow-inside-the-cyber-warfare-targeting-irans-critical-infrastructure
[71] NPR. (2021, October 27). A cyberattack paralyzed every gas station in Iran. https://www.npr.org/2021/10/27/1049566231/irans-president-says-cyberattack-was-meant-to-create-disorder-at-gas-pumps
[72] SCADAfence. (n.d.). Industrial cyber attack on Iranian steel companies explained. *Claroty Blog*. https://blog.scadafence.com/the-iran-steel-industry-cyber-attack-explained
[73]Malwarebytes. (2022, July 14). Predatory Sparrow massively disrupts steel factories while keeping workers safe. *Malwarebytes Blog*. https://www.malwarebytes.com/blog/news/2022/07/predatory-sparrow-massively-disrupts-steel-factories-while-keeping-workers-safe
[74] Kapko (2025).
[75] TechCrunch. (2025, June 17). Pro-Israel hacktivist group claims responsibility for alleged Iranian bank hack. https://techcrunch.com/2025/06/17/pro-israel-hacktivist-group-claims-responsibility-for-alleged-iranian-bank-hack/
[76] TRM Labs. (2025, June). Inside the Nobitex breach: What the leaked source code reveals about Iran's crypto infrastructure. https://www.trmlabs.com/resources/blog/inside-the-nobitex-breach-what-the-leaked-source-code-reveals-about-irans-crypto-infrastructure
[77] Wall Street Journal. (2025, June 29). How Israel-aligned hackers hobbled Iran's financial system. https://www.wsj.com/world/middle-east/how-israel-aligned-hackers-hobbled-irans-financial-system-fb1b0376
[78] METR. (2026, January 29). Time Horizon 1.1 [Blog post]. https://metr.org/blog/2026-1-29-time-horizon-1-1/
[79] Gartzke, E. (2013). The myth of cyberwar: Bringing war in cyberspace back down to earth. *International Security*, *38*(2), 41–73.
[80] Zetter, K. (2014). *Countdown to zero day: Stuxnet and the launch of the world's first digital weapon*. Crown.
[81] Trabish, H. K. (2025, February 12). Transformer supply bottleneck threatens power system stability as load grows. Utility Dive. https://www.utilitydive.com/news/electric-transformer-shortage-nrel-niac/738947
[82] United Nations. (n.d.). *Multilateral system*. https://www.un.org/en/global-issues/multilateral-system
[83] National Institute of Standards and Technology. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile* (NIST AI 600-1). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
[84] Megas, K., Cuthill, B., Snyder, J. N., Patrick, B., Khemani, I., Dotter, M., Garris, M., Zarei, M., & Schiro, N. (2025). *Cybersecurity AI profile: NIST community profile* (NIST IR 8596). National Institute of Standards and Technology.
[85] MITRE. (n.d.). *MITRE ATT&CK®*. Retrieved December 31, 2025, from https://attack.mitre.org/
[86] Chang, A., Saade, T., Mendapara, S., Swanda, A., & Garg, A. (2025). Cisco integrated AI security and safety framework report. *arXiv*. https://arxiv.org/abs/2512.12921
[87] Lin, J. W., Jones, E. K., Jasper, D. J., Ho, E. J. S., Wu, A., Yang, A. T., & Ho, D. E. (2025). Comparing AI agents to cybersecurity professionals in real-world penetration testing. *arXiv*. https://arxiv.org/abs/2512.09882
[88] Kouremetis, M., et al. (2025). OCCULT: Evaluating large language models for offensive cyber operation capabilities. *arXiv*. https://arxiv.org/abs/2502.15797
[89] Caldwell, S., Harley, M., Kouremetis, M., Abruzzo, V., & Pearce, W. (2025). PentestJudge: Judging agent behavior against operational requirements. *arXiv*. https://arxiv.org/abs/2508.02921
[90] Dawson, A., Mulla, R., Landers, N., & Caldwell, S. (2025). AIRTBench: Measuring autonomous AI red teaming capabilities in language models. *arXiv*. https://arxiv.org/abs/2506.14682
[91] Dreadnode. (2025). Evals: The foundation for autonomous offensive security [Blog post]. https://dreadnode.io/blog/evals-the-foundation-for-autonomous-offensive-security
[92] United Nations Secretary-General's High-Level Advisory Body on Artificial Intelligence. (2024). *Governing AI for humanity: Final report*. United Nations. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

[93] United Nations General Assembly. (2024). *Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development* (Resolution 78/265). United Nations. https://docs.un.org/en/A/res/78/265

[94] Cybersecurity and Infrastructure Security Agency. (2017, July 1). *Petya ransomware* (Alert TA17-181A). U.S. Department of Homeland Security. https://www.cisa.gov/news-events/alerts/2017/07/01/petya-ransomware

[95] United Nations Secretary-General's High-Level Advisory Body on Artificial Intelligence. (2024). *Governing AI for humanity: Final report*. United Nations. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

[96] Milmo, D., & Courea, E. (2025, February 11). US and UK refuse to sign Paris summit declaration on "inclusive" AI. *The Guardian*. https://www.theguardian.com/technology/2025/feb/11/us-uk-paris-ai-summit-artificial-intelligence-declaration

[97] Tasioulas, J. (2025, February 14). Expert comment: Paris AI summit misses opportunity for global AI governance. *University of Oxford*. https://www.ox.ac.uk/news/2025-02-14-expert-comment-paris-ai-summit-misses-opportunity-global-ai-governance

[98] Bengio, Y., Clare, S., Prunkl, C., Murray, M., Andriushchenko, M., Bucknall, B., Bommasani, R., Casper, S., Davidson, T., Douglas, R., Duvenaud, D., Fox, P., Gohar, U., Hadshar, R., Ho, A., Hu, T., Jones, C., Kapoor, S., Kasirzadeh, A., . . . Mindermann, S. (2026). *International AI Safety Report 2026* (DSIT 2026/001). Department for Science, Innovation and Technology. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

[99] UK Government. (2023, November 1). *The Bletchley Declaration by countries attending the AI Safety Summit, 1–2 November 2023*. GOV.UK. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

[100] Federal Communications Commission. (2025). U.S. Cyber Trust Mark. https://www.fcc.gov/CyberTrustMark

[101] Office of the Federal Register. (2024, October 15). Cybersecurity Maturity Model Certification (CMMC) Program. Federal Register, 89(199), 83092-83237. https://www.govinfo.gov/app/details/FR-2024-10-15/2024-22905

[102] U.S. General Services Administration. (2023). FedRAMP cloud service provider authorization playbook. https://www.fedramp.gov/

[103] Agentic AI Foundation (n.d.). Agentic AI Foundation. https://www.agenticai.foundation/

[104] Lloyd's. (2022, August 16). State backed cyber-attack exclusions (Market Bulletin Y5381). https://assets.lloyds.com/media/35926dc8-c885-497b-aed8-6d2f87c1415d/Y5381%20Market%20Bulletin%20-%20Cyber-attack%20exclusions.pdf

[105] Lloyd's. (2024, May 14). *State-backed cyber-attack wordings* (Market Bulletin Ref: Y5433). https://lmalloyds.com/wp-content/uploads/2025/06/Y5433.pdf

[106] Bueker, J. P., Hallward-Driemeier, D., Hoey, L. G., Kossak, A. D., Lampert, M. B., Levy, J. S., & O'Connor, A. (2026, January 20). False Claims Act insights: Key takeaways from DOJ's fiscal year 2025 cases & recoveries. Ropes & Gray. https://www.ropesgray.com/en/insights/alerts/2026/01/false-claims-act-insights-key-takeaways-from-dojs-fiscal-year-2025-cases-recoveries

[107] Major Fraud Act of 1988, 18 U.S.C. § 1031 (2018).