THE CYBER DEFENSE REVIEW
In press, 28 pages
https://doi.org/10.55682/cdr/gcai-agda

RESEARCH ARTICLE

Grand Challenges in Agentic AI for Cyber Operations: A Research Agenda

Kyle Dotterrer, Allyson I. Hauptman, Blane P. Richoux, Matthew L. Corbett*

Army Cyber Institute, West Point, NY, USA

Agentic AI systems are reshaping cyber operations at a pace that outstrips the mechanisms needed to deploy them responsibly. The first documented autonomous cyber attack, in September 2025, demonstrated that the technology has crossed the threshold from research capability to operational threat, yet the technical robustness, human-AI trust, and governance frameworks required for responsible adoption remain underdeveloped. This paper argues that the resulting gaps constitute a control deficit manifested through a set of grand challenges across four reinforcing dimensions: technical limitations and vulnerabilities, the trust deficit between operators and AI agents, insufficient governance, and dual-use escalation risks. Through a structured expert analysis that draws on operational, technical, human, and policy perspectives, we characterize the interactions among these dimensions. We then propose a research agenda to help the cyber operations community address these grand challenges in a coordinated manner, accounting for both the dependencies among dimensions and the operational urgency posed by the adversary's adoption of the same technology. In doing so, the paper calls on researchers, practitioners, and policy-makers to collectively shape the responsible integration of agentic AI into cyber operations without ceding the advantages it provides.

Keywords: cyber warfare, Artificial Intelligence (AI), agentic AI, cyber operations, autonomy, control deficit, national security, research agenda, grand challenges

* Corresponding author: matthew.corbett@westpoint.edu

Disclaimer: *The views expressed in this work are those of the author(s) and do not reflect the official policy or position of their employer(s), the U.S. Military Academy, the Department of War, the U.S. Government, or any subdivisions thereof. 2026. This is a work of the U.S. Government and is not subject to copyright protection in the United States. Foreign copyrights may apply.*

1 INTRODUCTION

Autonomous AI agents are reshaping the tempo and character of activity in cyberspace. Once confined to research labs and controlled exercises, agentic systems now plan, sequence, and execute complex tasks across the defensive and offensive spectrum (AISI 2025; Anthropic 2025a; Fink 2025; Friar and Payne 2025). This acceleration promises substantive gains in detection, response, and resilience, but it also amplifies unpredictability, complicates oversight, and lowers barriers to malicious misuse (Anthropic 2025a). As organizations experiment with deployment in high-stakes environments, they face a widening gap between capability and control: tools advance faster than the mechanisms for understanding, governing, and integrating them. This article examines that gap, characterizing the dimensions along which control has failed to keep pace with capability, identifying the dynamics among these dimensions, and proposing a research agenda to close the resulting deficit without ceding operational advantage to adversaries.

1.1 The First Autonomous Cyber Attack

In September 2025, the hyperscale AI company *Anthropic* detected misuse of its flagship LLM, Claude, in a first-of-its-kind cyberattack (Anthropic 2025b). In a detailed report, Anthropic highlighted the incident as the first known use of an LLM for a fully automated, long-duration, self-directed cyberattack with limited human intervention. This attack was analyzed and exploited targeted networks at a scale beyond that attainable with direct human decision-making and oversight, representing a watershed moment in the history of cyberspace conflict. Despite failings in Claude's reasoning and methods, the attack successfully compromised multiple networks simultaneously using open-source penetration-testing tools, which the LLM-based agent directed with near independence.

The September attack is a bellwether for the future of military and civilian operations in cyberspace.[1] Autonomous offensive operations are now fully within the realm of possibility, and the implications for the cyber domain are vast and pressing (Manning 2025). Network defenders must now react to attacks that will progress at the speed of LLM token generation rather than human direction. In anticipation of these adversary capabilities, cyberspace defenders are automating network defense (Friar and Payne 2025). Autonomous agents capable of detecting intrusions and deploying countermeasures in real time may soon match the speed and autonomy of offensive agents.

1.2 The Control Deficit

This turning point reveals an underlying, systemic issue with autonomous cyber systems. The swift deployment of agentic AI in cyberspace has created concurrent shortfalls in technical robustness, human-AI trust, and governance — deficiencies that together constitute what we term the *control deficit*: an expanding gap between the capabilities of agentic AI systems and the operational, technical, human, and governance mechanisms to employ them responsibly. Autonomous agents remain constrained by technical limitations that prevent them from scaling beyond controlled environments (Oesch et al. 2024); the security risks introduced by agentic architectures and protocols such as MCP remain under active investigation (Gabarda 2025); trust and explainability challenges in human-AI teaming are well documented (Duan et al. 2025; Hauptman, Mallick, et al. 2025); and frontier safety frameworks

1. For purposes of this article, references to *cyberspace operations* align generally with the doctrinal usage in Joint Publication (JP) 3-12. (Joint Chiefs of Staff 2022)

continue to evolve to address escalation risks (Forum 2025). Yet these efforts proceed largely in parallel, without a unified account of the dynamics that govern their interactions. Our analysis demonstrates that these conditions interact as a self-reinforcing cycle, making it difficult to resolve any one dimension without simultaneously addressing the others. The terms “agentic” or “agents” have recently been used to describe Large Language Model (LLM)-driven, autonomous (or partially autonomous) systems capable of action without direct human oversight (Stackpole 2026). While the topic of autonomy in cyberspace is broader than agentic AI, our focus on this technology reflects its predominance among autonomous systems in the cyberspace domain.

In this paper, we draw on technical, operational, and policy perspectives through collaborative analysis to address the control deficit directly. We characterize the deficit as a structured problem space comprising four reinforcing dimensions and present a research agenda to close its gaps for the community of researchers, practitioners, and policy-makers tasked with responsibly integrating agentic AI into cyber operations. The following section outlines the technical context for this analysis by surveying the evolution of autonomous cyber operations, from their origins to the agentic systems reshaping today’s landscape.

2 STATE OF THE ART

2.1 A Decade of Development

In 2016, the state of the art in autonomous cybersecurity was on display at the Paris Las Vegas Hotel & Conference Center in the Mojave Desert. Alongside the annual DEF CON hacker convention, the venue hosted the Defense Advanced Research Projects Agency (DARPA) Cyber Grand Challenge (CGC) finals. In this challenge, competitors deployed what they termed "Cyber Reasoning Systems" (CRS) to autonomously detect, evaluate, and patch software vulnerabilities in real time, as well as exploit any vulnerabilities they discovered in competitors’ systems (DARPA 2016). To achieve this, competing project teams developed novel automation methods for computer security tasks, including vulnerability identification, binary patching, exploit development, and implementation of security strategies. Such methods allowed a CRS to perform select cyberspace offense and defense tasks without human supervision (Shellphish 2016; Shoshitaishvili et al. 2018; Stephens et al. 2016).

Although technically impressive, the CRS systems demonstrated important technical limitations. Built on program analysis techniques—principally fuzzing, symbolic execution, and binary patching—they were limited to predefined sets of binary formats and vulnerability classes during competition (DARPA 2017). While performing well within those boundaries, their methods were tightly coupled to the competition’s specifications and did not generalize to heterogeneous software, diverse network topologies, or unpredictable environmental conditions (Song and Alves-Foss 2016). Although the CRSs at the 2016 CGC did not immediately transform cybersecurity practice, they foreshadowed a trajectory whose significance is clearer in retrospect.

What has changed over the intervening decade is not the underlying ambition, but the maturity of the technologies capable of realizing it. Autonomous cyber operations are now a more practical reality with significant implications for the security posture of operational systems, organizations, and enterprises. Much of that progress has stemmed from the rise of *agentic AI*. While the term *agent* entered routine use in the AI field as early as the mid-1990s (Russel and Norvig 2020), only recent

advances in reinforcement learning and large language models have made agentic systems both powerful and practical. Reinforcement learning (RL), particularly deep RL, provides a framework for autonomous, goal-directed learning, allowing agents to develop effective strategies in complex, dynamic environments through trial-and-error feedback. LLMs complement this by providing extensive world knowledge acquired through internet-scale pretraining, enabling agents to reason about context, generate code, and communicate in natural language. Together, these paradigms make it increasingly feasible for agents to achieve specific objectives with minimal supervision, including vulnerability discovery, exploit development, and end-to-end attack and defense automation.

2.2 Modern Agents in Cyber Operations

Agentic AI systems are transforming how software vulnerabilities are analyzed and discovered. Systems like IRIS (Li, Dutta, and Naik 2025) combine static analysis with automatic specification generation to discover security vulnerabilities in source code. By combining traditional program analysis techniques, such as those demonstrated by the CRSs at the 2016 CGC, with the contextual reasoning and generative capabilities of LLMs, these systems scale to larger codebases and require less human oversight. As a result, these neuro-symbolic approaches can identify vulnerabilities at a scale that previously would have required multiple human experts.

Advances in agentic methods and their underlying language models are also enabling new paradigms of vulnerability analysis. OpenAI's agentic security researcher, *Aardvark*, for example, leverages source code understanding and tool use to identify program vulnerabilities. Unlike traditional program analysis techniques like fuzzing and symbolic execution, Aardvark reads and analyzes code as a human security researcher would. The agent employs LLM-powered reasoning to review source, scan commits for changes, identify newly introduced vulnerabilities, validate them in sandboxed environments, and generate patches (OpenAI 2025b). This capability, along with others like it, supports direct integration with developer workflows, enabling many more individuals and teams to contribute to internal vulnerability research, identification, and remediation.

Beyond vulnerability discovery, LLM-based agents are demonstrating increasing proficiency in developing working exploits. Early results showed that single GPT-4 agents can autonomously exploit 87% of one-day vulnerabilities when provided with CVE descriptions (Fang et al. 2024). These same agents, however, performed poorly against zero-day vulnerabilities — real-world flaws of which the agent has no *a priori* knowledge — exposing a ceiling on what single-agent architectures can achieve. Recent approaches overcome this challenge by decomposing exploit development into a multi-task, cooperative activity and aligning specialized agents to each task, enabling more effective exploitation of previously unknown vulnerabilities (0-days) (Zhu et al. 2025).

Modern cyber operations are employing AI agents in both defense and offense. On the defensive side, autonomous cyber defense (ACD) agents identify, analyze, and counteract threats in real time with minimal human intervention (Castro et al. 2025). Most current approaches rely on reinforcement learning, training agents through simulated engagements in environments that model operational network conditions. RL-based agents excel at rapid and tactical decision-making, such as selecting containment actions or adjusting firewall rules under time pressure, but struggle with tasks that require contextual reasoning or generalization to unfamiliar network configurations (Purves et al. 2024; King, Bowman, and Huang 2025). On the offensive side, agentic systems demonstrate an

ability to orchestrate multi-stage attacks across complex network topologies. LLM-based agents can autonomously replicate breach patterns, including initial exploitation, malware installation, lateral movement, and data exfiltration activities (Singer et al. 2025b). Initially, such capabilities required specialized offensive frameworks to guide LLM-based agents through multi-stage operations (Singer et al. 2025a). Increasingly, frontier LLMs are proving effective at multi-host network attack without any additional infrastructure (Anthropic 2026b).

Agentic AI has progressed from narrow, task-specific automation to systems capable of executing complete cyber operation workflows with minimal human oversight. The capability trajectory from the constrained CRSs of 2016 to agents adept at discovering vulnerabilities, developing working exploits, and orchestrating end-to-end attacks has been rapid. A critical asymmetry has emerged in this progression: offensive automation can be deployed by individual actors with minimal infrastructure, while defensive automation requires extensive integration, validation, and organizational coordination. Whether the field can close this asymmetry depends on its ability to address the gaps that currently prevent responsible adoption—the subject of the remainder of this paper.

3 RESEARCH OBJECTIVES AND METHODOLOGY

3.1 Objectives and Scope of Research

The trajectory from the 2016 Cyber Grand Challenge to the autonomous attacks of 2025 demonstrates that agentic capability in cyberspace is no longer a theoretical concern. It is an operational reality. The rapid operationalization of agentic AI in cyberspace has created simultaneous gaps in technical robustness, human-AI trust, and governance that together impede the responsible adoption of this technology. Each of these gaps is recognized in existing literature, but the field lacks a unified characterization of their interactions and an agenda for addressing them under the time pressure imposed by adversary adoption. This paper aims to provide both. To this end, the research team focused its analysis on the following research question:

What critical gaps impede the responsible and effective adoption of agentic AI in cyber operations, and what research priorities should guide efforts to close them?

This research assumes that agentic systems will become integral to daily offensive and defensive cyber operations in the near future. To characterize this problem space comprehensively, the research team scoped its analysis to agentic AI systems that could realistically be integrated into cyber operations within the next decade, enabling a forward-looking analysis while remaining grounded in the current state of the art. Within this scope, the team structured its analysis along four dimensions, each addressed by a guiding sub-question:

(1) **Operational**: What capability gaps prevent current agentic AI systems from scaling to real-world cyber operations, both offensive and defensive?
(2) **Technical**: What new risks and attack surfaces do agentic architectures introduce to the cyber landscape?
(3) **Human**: How do trust and explainability challenges in agentic AI constrain its adoption in cyber operations?
(4) **Governance**: What legal and policy gaps impede the responsible governance of agentic AI in cyber operations?

3.2 Methodology

Research Team Composition and Positionality. Characterizing a problem space that spans technical, operational, human, and governance dimensions requires a research team with corresponding breadth of expertise. The research team involved in this analysis is based at a U.S. military research institute and holds degrees in computer science, data science, cybersecurity, information security, and human-centered computing, with real-world experience in AI and cyber operations. Additionally, the team's broad research portfolio provided the study with a solid foundation for understanding the specific challenges explored in this analysis.

Accordingly, the perspectives and priorities reflected in this work are shaped in part by the operational realities, governance structures, and strategic considerations of the U.S. military and national security context, particularly with respect to cyber operations and emerging AI capabilities, while also drawing on academic traditions and interdisciplinary perspectives developed outside military institutions.

Analysis of the Problem Space. The research team sought to identify challenges and opportunities across the problem space through iterative group discussions, in which experts reviewed relevant literature, their research findings, and areas for future research in their fields. These discussions took place as weekly meetings over a three-month period and were organized into three phases: subject-area reviews, challenge identification, and research-agenda creation.

For the review process, the research team first considered sources from both academia and industry. Academic sources were used to assess the state of the art, with emphasis on evaluations of technical capabilities and current trends in practical applications. As industrial entities are often at the forefront of this technology, particularly those published by frontier model providers, their work was used to establish near-term projections and identify open societal-level challenges. Finally, to ground the work in the operational realities of offensive and defensive cyber operations, the research team leveraged its own experience and that of a select group of anonymous subject-matter experts.

Elicitation of the Grand Challenges. The selection phase was an iterative, bottom-up process in which the researchers first presented *all* the challenges they identified in their topic area to the rest of the team and collectively selected those most closely related to the research questions.

Challenges were retained for inclusion when they satisfied a set of criteria associated with grand challenges in emerging technological domains (Elvitigala et al. 2024). Specifically, the research team prioritized challenges that were highly salient to agentic AI in cyber operations, and whose unresolved status would significantly constrain the military utility, scalability, trustworthiness, or responsible governance of these systems. The selected challenges were required to involve substantial technical, operational, human, legal, or strategic complexity, such that they could not be resolved through incremental improvements or isolated technical advances alone. The team also focused on challenges that could realistically be addressed through sustained interdisciplinary efforts over the coming decade, while requiring coordination across multiple stakeholder communities or interdisciplinary academic collaboration beyond the scope of a single method, institution, or research group. Finally, the team prioritized challenges whose resolution would be likely to produce transformative effects on cyber operational practices, governance frameworks, strategic stability, or the broader role of agentic

AI in cyber operations. Thus, the challenges and possible solutions presented in the remainder of this study reflect the team's view of the greatest, most imminent, and most tractable challenges to agentic AI in cyber operations.

Definition of a Coordinated Research and Governance Agenda. The final phase of the analysis involved developing a coordinated agenda to guide future research, operational development, and governance efforts related to agentic AI in cyber operations. For this phase, the research team examined how the identified operational, technical, human, and governance challenges interacted across the broader problem space, with particular attention to the reinforcing dynamics across dimensions that complicate narrowly scoped or isolated interventions. Through iterative synthesis and discussion, the researchers explored different ways to organize and prioritize coordinated lines of effort, distinguishing challenges that required foundational progress from those that could proceed in parallel under operational time pressure. This process informed the systems-level structure of the agenda presented in Section 5.

Noteworthily, the analysis is interpretive and synthetic in nature. The resulting agenda should therefore be understood as an expert-driven, forward-looking effort to characterize a rapidly evolving problem space, identify key interdependencies among challenges, and propose coordinated directions for future research, operational development, and governance.

4 GRAND CHALLENGES: THE CONTROL DEFICIT IN AGENTIC CYBER OPERATIONS

As described in Section 1, the *control deficit* is a widening gap between what agentic systems can do and our ability to make them trustworthy, secure, and governable in operational settings. This deficit manifests across four dimensions that we identify and characterize in this section: technical limitations and vulnerabilities that prevent current agents from scaling beyond controlled environments and present new attack surfaces, a trust deficit rooted in the inherent opacity of AI decision-making, insufficient governance frameworks for autonomous cyber operations, and dual-use[2] escalation risks that the current policy landscape is ill-equipped to manage. Critically, these dimensions are not independent — they interact in ways that make the aggregate problem harder to solve than any single gap in isolation. We examine each dimension in turn before analyzing the way in which they reinforce one another.

4.1 Capability Gaps and Vulnerabilities Limit Operational Readiness

Recent demonstrations of agent-driven autonomous cyber operations have generated considerable excitement, but current capabilities fall short of the demands of real-world operations. Both offensive and defensive applications face distinct technical limitations that must be overcome before these systems can scale beyond controlled environments.

4.1.1 Autonomous Cyber Offense. Agents for offensive cyber operations have advanced rapidly, crossing the threshold from proof-of-concept to operational relevance. Despite impressive demonstrations, current systems still exhibit critical capability gaps that prevent them from scaling to support sophisticated autonomous offensive operations. Recent evaluations across multiple research groups have begun to characterize these gaps with increasing precision.

2. In the literature on frontier AI and cyber capability risk, *dual-use* and *dual-purpose* commonly refer to capabilities with both offensive and defensive applications. Our use of those terms in this article falls within that context.

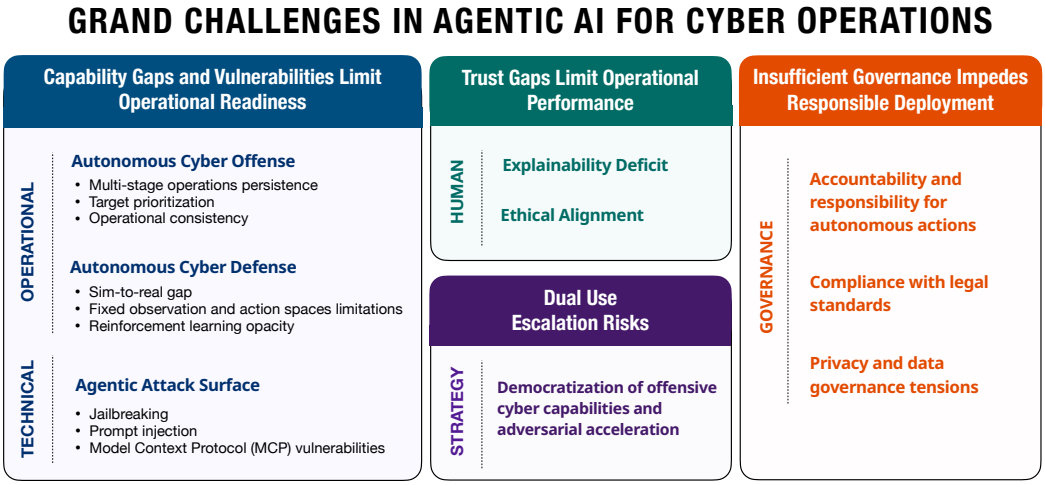

Figure 1. Grand challenges in Agentic AI for cyber operations

The most consequential limitation is an inability to sustain coherent, multi-stage operations. Although current agents can perform individual offensive tasks, they struggle to combine these into adaptive, goal-directed chains that characterize real-world attacks. Frontier models such as GPT-5.2 Codex still achieve zero-percent success rates on CyScenarioBench—a benchmark designed to measure multi-stage cyber operations under realistic constraints (OpenAI 2025a). When assisted by sophisticated cyber attack 'harnesses', LLM-based agents achieve only partial success in multi-stage attack scenarios, frequently stopping after a single objective rather than continuing to explore additional attack paths (Singer et al. 2025b). Even when the supporting infrastructure identifies additional targets, agents often lack the persistence to pursue them (Lin et al. 2025).

Beyond persistence, current agents exhibit a gap between execution and identification. When directed toward a specific vulnerability, agents can often exploit it; the bottleneck is recognizing what to investigate in the first place. When confronted with a large, complex network environment representative of real-world scenarios, even the most competent agents can be distracted by minor misconfigurations when critical remote code execution vulnerabilities are nearby (Lin et al. 2025). Such distinctions, trivially apparent to attackers or penetration testers, remain invisible to agents that cannot distinguish high-value information from noise.

Consistency is another critical shortcoming. In operational scenarios, one often does not have multiple opportunities to execute certain stages of a multi-stage cyber attack; active defensive measures, such as network and host monitoring, preclude such an approach. This reality highlights the distinction between *capability* and *reliability*, where even the most capable models rely on multiple trials to achieve success in multi-stage attacks (Singer et al. 2025b) and continue to exhibit "a consistent pattern of mistakes" on current benchmarks (OpenAI 2025a).

In military cyber contexts, these limitations are compounded by operational demands that no current benchmark attempts to measure: weighing outcomes against attribution risk, managing operational security across extended campaigns, and exercising judgment about tool exposure. Until agents

demonstrate consistent, multi-stage operational capability against realistic targets under adversarial pressure, fully autonomous offensive cyber operations against sophisticated adversaries will remain out of reach.

4.1.2 Autonomous Cyber Defense. Agents for defensive cyber operations face more fundamental technical barriers to operational deployment than their offensive counterparts. Where offensive agents can demonstrate value through narrow task execution even when they cannot sustain full campaigns, defensive agents must operate continuously across complex, dynamic networks; this requirement exposes the fragility of current approaches.

The most pressing limitation is the *sim-to-real gap*: the distance between the training environments where ACD agents learn and the production networks where they must perform (Vyas et al. 2023). Popular platforms like CybORG (Standen et al. 2021) provide a foundation for research but do not yet capture the complexity of operational environments. Observation spaces are artificially simplified abstractions that assume complete information about network state—an assumption that does not hold in production environments, where defenders operate with partial, noisy, and often delayed visibility (Oesch et al. 2024). They also frequently force agents to re-learn functionality already provided by existing defensive tools such as endpoint detection and response (EDR) systems and host-based intrusion detection systems (HIDS), rather than learning to synthesize these tools as a human analyst would (Oesch et al. 2024).

Beyond environmental fidelity, ACD agents face generalization challenges rooted in the architectural assumptions of deep reinforcement learning. Deep RL agents are typically designed around fixed-size observation and action spaces (Mnih et al. 2015), which impose significant barriers to portability and adaptability. A fixed observation space means that an agent trained on one network topology cannot readily transfer to a network of a different size or structure—each new environment requires retraining. A fixed action space means that incorporating new defensive actions in response to emerging threats or new intelligence requires modifying the agent's architecture, rather than simply updating its knowledge. These rigidities are compounded by the volume of training data required by deep RL agents (Oesch et al. 2024), as retraining is not merely an architectural inconvenience but a resource-intensive process that limits how quickly agents can be adapted to new operational contexts. These fixed-space limitations represent fundamental barriers to producing agents that generalize across the diverse and evolving networks they would encounter in operational deployment.

Finally, the reinforcement learning paradigm underlying most current ACD agents poses an explainability challenge with direct consequences for operational adoption. Deep RL agents trained with popular algorithms can learn effective cyber defense policies, but their decision-making is opaque because the learned policy is encoded in neural network weights that are difficult to interpret. This opacity is a barrier in any AI application, but it is particularly acute in defensive cyber operations, where the consequences of an incorrect action may include disrupting friendly systems or creating gaps that adversaries can exploit. In such settings, human operators must be able to understand why an agent took a given action, assess whether intervention is appropriate, and maintain the level of accountability required by existing policy frameworks. Without this understanding, the trust necessary for meaningful human-AI teaming in defensive operations cannot be established.

4.1.3 Agents Introduce a New Attack Surface. Like all software, the language models that power LLM-based agents are susceptible to security vulnerabilities. Attackers conduct jailbreaks to induce unintended behavior, thereby bypassing the internal security measures implemented by the model's developers. Jailbreaks take many forms, including clever disguises of prompt intent (Wei, Haghtalab, and Steinhardt 2023), the use of alternative languages (Wei, Haghtalab, and Steinhardt 2023), and universal transferable suffixes embedded in textual or visual input (Qi et al. 2023; Zou et al. 2023). In LLM-based agents, such vulnerabilities allow an attacker to steer the agent's behavior beyond established guardrails, violating the security of the systems the agent is intended to protect.

While jailbreaks are undesirable, prompt injection remains the foremost unsolved problem in frontier model security (Stuckey 2025). In this class of LLM exploitation, attackers influence a model's responses through carefully crafted intermediate content, enabling them to carry out a variety of malicious acts, from credential theft to data exfiltration (Greshake et al. 2023; Rehberger 2023). Unlike traditional computing systems, where hardware-enforced boundaries separate executable code from data, LLMs process all inputs as a unified stream of tokens (Radford et al. 2019). This architectural reality implies there is no principled mechanism by which a model might distinguish between commands it should execute and text it should merely process. A 2025 paper co-authored by researchers from OpenAI, Anthropic, and Google DeepMind underscored the severity of this challenge: by systematically tuning general optimization techniques, the authors bypassed 12 recently published defenses, achieving attack success rates exceeding 90% (Nasr et al. 2025).

LLM-based agents derive their world knowledge and autonomous behavior from pretrained foundation language models. To empower them to use this knowledge to achieve objectives on our behalf, we provide them with a flexible means to interact with the world, gather new information, and take actions that influence external systems. An emerging interface protocol intended to standardize this interaction is the Model Context Protocol (MCP). MCP provides a universal means for LLM-based agents to interact with any external system, including data sources, tools, and workflows (Anthropic 2024). It also provides attackers with a new means to compromise agentic systems by gaining control of MCP servers, a critical component of the MCP architecture, thereby enabling them to manipulate data flow and tool calls from client applications. MCP servers are susceptible to attacks, such as unauthorized command execution, that exploit improperly sanitized input. Once an attacker controls an MCP server, they can subsequently compromise the behavior of any agents that make use of it through techniques such as tool injection, using tool calls to gather confidential data, or instructing the agent to take other undesirable actions on the attacker's behalf (Gabarda 2025).

These model-level and system-level vulnerabilities compound the challenges described above: agents that cannot yet perform reliably in real-world operations are simultaneously difficult to secure against adversarial manipulation. Together, these technical shortcomings have direct consequences for the human operators expected to work alongside these systems, as described in the next section.

4.2 Trust Gaps Limit Operational Performance

As humans come to rely on and collaborate with AI agents, the degree to which they trust and understand those agents becomes increasingly important. Human-AI teaming research demonstrates that both trust in an AI agent and human understanding of its decisions heavily influence overall team performance (McNeese et al. 2021). Two factors pose particular threats to this trust: the inherent

opacity of AI decision-making and the risk that AI agents may behave in ways perceived as unethical. Addressing these challenges is critical to capturing the full benefits of agentic AI in cyber operations.

A recent review of trust in human-AI teams found that specific human-AI interaction factors, particularly explainability, have the greatest impact on the generation and maintenance of trust (Duan et al. 2025). Modern agents are based on machine learning and typically employ neural networks as their underlying architecture. These networks encode the patterns they learn as complex mathematical objects, making them inherently difficult to understand and creating a natural barrier to native explainability (Şahin, Arslan, and Özdemir 2025). Without the ability to explain an agent's actions, operators lack the understanding necessary to predict future behavior and forfeit the opportunity to improve the agent when things go wrong. Both of these factors degrade trust. This explainability gap is pervasive in the automated systems deployed today, both for offense and defense. Left unaddressed, it will continue to limit our ability to leverage agentic AI to its full potential.

Beyond explainability, ethics play a prominent role in a human's decision to use, trust, and collaborate with an AI agent. Research demonstrates that programming an AI system with a concept of its team's ethical code significantly increases human collaborators' trust and acceptance of the agent (Hauptman, Schelble, et al. 2025). However, the inverse also applies: when an AI agent makes what is perceived as an unethical decision, it shatters human confidence, undermining the trust in the agent itself and the entire organization in which it is employed (B. Schelble et al. 2023). If an organization continues to use an AI agent that its employees perceive as unethical, it can quickly lead to skepticism and paranoia among the workforce. Ensuring that AI agents act in accordance with human ethical expectations is therefore not optional; it is a necessary condition for the trust that underpins effective human-AI teaming.

The trust challenges described here have implications beyond individual team performance. Organizations that cannot establish sufficient trust in agentic systems will inevitably delay adoption. Governance frameworks require operational experience to be well-calibrated, yet that experience cannot accumulate without deployment.

4.3 Insufficient Governance Impedes Responsible Deployment

Autonomous AI systems pose unique challenges for the existing legal and policy frameworks designed to govern cyber operations. AI governance—the set of regulations, methods, procedures, and technical mechanisms used to ensure AI systems align with organizational goals and societal values (Batool, Zowghi, and Bano 2025)—has not kept pace with the rapid operationalization of agentic systems. Existing frameworks were designed to govern human decision-makers or, at most, narrowly scoped automated tools operating under direct human supervision. While international bodies have provided commentary on applying existing frameworks to new technologies in cyberspace (Schmitt 2013), and nations have taken initial steps to establish their positions on autonomous weapons (Hicks 2023), no organization has specifically addressed the novel challenges autonomous AI poses.

As a result, current approaches do not address the novel challenges introduced by agents that plan, adapt, and act with increasing independence: attributing responsibility for autonomous actions, ensuring proportionality at rapid speeds, and maintaining compliance with legal standards designed for human cognition. Without governance frameworks calibrated to these realities, organizations face

a choice between deploying systems they cannot adequately oversee and delaying adoption while adversaries face no comparable constraint.

The law of armed conflict (LOAC) illustrates both the limitations and the latent promise of existing governance frameworks when applied to autonomous agents. The four bedrock LOAC principles—distinction, proportionality, unnecessary suffering, and military necessity—are designed to limit the horrors of war and rely upon the judicious application of human reason (Council 1978). Autonomous agents that operate without direct human oversight fundamentally challenge this reliance: if no human is reasoning about these principles and their balance, how is compliance assured? Yet the relationship is not purely adversarial. Research on AI agents in high-risk contexts indicates that a human-in-the-loop is not always necessary for AI to operate in the spirit of LOAC (Hauptman, Schelble, et al. 2025). Agents may, in fact, reduce errors caused by human cognitive impairment under stress, improving compliance in domains where human judgment is most fallible. LOAC is representative of a broader class of governance frameworks—rules of engagement, operational authorities, oversight requirements—that were designed around the assumption of human agency and will need to be revisited as non-human agents assume increasingly autonomous roles in operations.

At the national policy level, there is a pronounced gap between strategic direction and operative frameworks. In the U.S., the previous and current administrations have used executive orders to guide AI development. These include those aimed at strengthening cybersecurity that included provisions for accelerating AI deployment to defend critical infrastructure (Biden 2025), and directed reviews of government policies that may hinder AI development and deployment (Trump 2025). While these orders encouraged enhanced cybersecurity and innovative AI programs, they did not produce new policies, laws, or authorities for the use of AI in cyber operations.

Data privacy exemplifies the kind of unresolved governance question that arises when policy lags behind capability. The performance of AI systems generally improves with the volume of data on which they are trained and with which they operate (Kaplan et al. 2020), creating tension between operational effectiveness and privacy protection (Mirishli 2024). In practice, this tension manifests along two axes. First, how much data should an agent collect from the operators alongside whom it works, and what consent should those individuals have over that collection? Second, in operational environments that include civilian infrastructure, what protections must govern the data an agent senses from its surroundings? These questions have no current answers in existing domestic or international regulation, and they highlight an area where the objectives of capability developers and policy-makers are in direct tension.

The absence of adequate governance frameworks is particularly consequential given the dual-use nature of AI in cyber operations. Without structured mechanisms to differentiate legitimate use from malicious exploitation, advances in agentic capability flow to adversaries as readily as to defenders.

4.4 Dual-Use Introduces Escalation Risks

Alongside weapons of mass destruction, cybersecurity is among the highest-risk domains for foundation models and AI agents (OpenAI 2023). The properties of cyberspace—its interconnectedness, the speed at which operations unfold, and its potential for automation—make cyber operations particularly amenable to rapid AI-induced acceleration. As early as 2024, members of the Frontier Model

Forum, including leading AI developers like OpenAI and Anthropic, identified cybersecurity applications as a significant source of risk resulting from advances in language model technology (Forum 2024; OpenAI 2023; Anthropic 2025d). These AI providers, along with national agencies like the US Center for AI Standards and Innovation (CAISI) and UK AI Security Institute (AISI), are concerned about an increase in the volume and impact of cyber operations as a result of agent capabilities (Folkerts et al. 2026; Lutnick 2025).

The underlying issue driving these risk concerns is AI's inherently dual-use nature in cybersecurity. Technological breakthroughs intended to enhance cyber defense often simultaneously enable offensive operations, thereby creating opportunities for actors who might otherwise be deterred by the technical expertise required to conduct attacks. Members of the Frontier Model Forum and those sympathetic to their claims are concerned about the democratization of sophisticated offensive cyber capabilities, which naturally results from rapid advances amid the dual-use dilemma. In their *Preparedness Framework*, OpenAI speculates on the implications of this scenario, observing that "removing bottlenecks limiting malicious cyber activity may upset the current cyberoffense-cyberdefense balance by significantly automating and scaling the volume of existing cyberattacks" (OpenAI 2025c). Other security experts are similarly apprehensive about the future of cybersecurity, envisioning "a singularity event for cyber attackers" after which the status quo that has reigned for so long is permanently upended (Adkins, Evron, and Schneier 2025).

4.5 The Reinforcing Structure of the Control Deficit

The four dimensions characterized above interact in ways that create a reinforcing structure, in which progress in any single dimension is constrained by the state of the others. The most direct chain of reinforcement begins with the technical foundations. The capability limitations and security vulnerabilities described in Sections 4.1 and 4.1.3 are the primary drivers of the trust deficit. Operators cannot trust systems whose decisions they cannot interpret, whose failure modes they cannot predict, and whose architectural components are susceptible to adversarial manipulation. When an agent's reasoning is opaque and its security posture is uncertain, the explainability and ethical alignment mechanisms on which trust depends lack a reliable foundation.

The trust deficit, in turn, constrains governance. Organizations that cannot establish sufficient confidence in agentic systems will delay their deployment, and governance frameworks require operational experience to be well-calibrated. Policies developed without the empirical grounding of real-world deployment risks can be either too restrictive, stifling adoption and ceding advantage to less risk-averse adversaries, or too permissive, failing to mitigate the risks they are designed to address. The current state of affairs reflects exactly this dynamic: executive orders direct the development and deployment of AI for cyber operations (Biden 2025), yet the policy frameworks within which developers should design and field these systems remain largely undefined. This vacuum is not merely a matter of legislative pace—it is a structural consequence of the fact that the trust necessary to drive adoption and the adoption necessary to inform policy have not yet materialized.

The governance vacuum is especially consequential given AI's dual-use nature in Cyber. Advances in agentic AI flow to adversaries as readily as to defenders. Adversary adoption, in turn, compresses the timeline for solving the technical and trust problems that initiated the cycle. Defenders face mounting

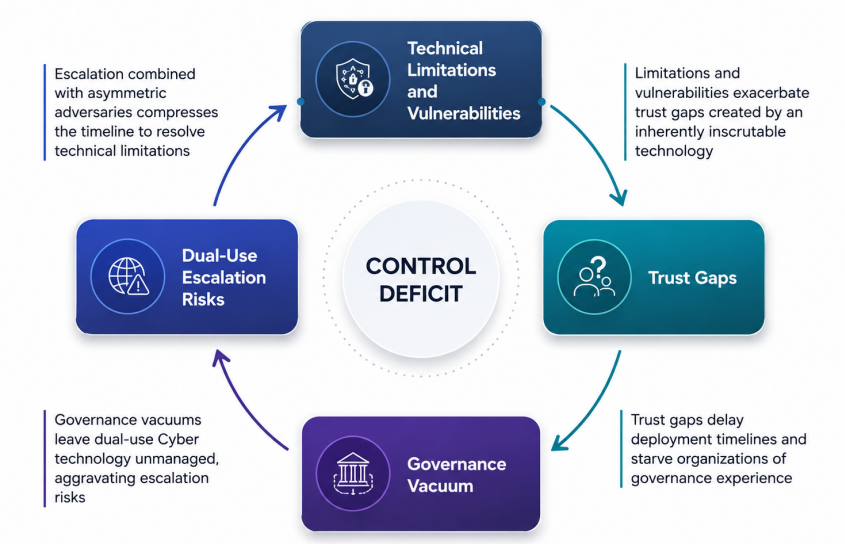

Figure 2. The control deficit as a self-reinforcing structure. Technical limitations and new attack surfaces, trust gaps, governance vacuums, and escalation risks form a self-reinforcing loop.

pressure to deploy agentic systems despite unresolved limitations, because the alternative cedes the initiative to opponents who face no comparable constraints on adoption.

This self-reinforcing property, summarized in Figure 2, is the central structural feature of the control deficit. Technical shortcomings inhibit trust; insufficient trust stalls adoption and starves governance; inadequate governance leaves dual-use unmanaged; adversary exploitation of dual-use capabilities heightens the urgency of resolving the technical shortcomings that began the cycle. Recognizing this structure is essential because it implies that narrowly scoped solutions will be insufficient if pursued in isolation. The reinforcing dynamics ensure that unaddressed dimensions will undermine progress on the others. This characterization of the control deficit as a structured, reinforcing problem space, rather than as a collection of independent challenges, motivates the research agenda presented in the following section. There, we propose two coordinated lines of effort to address these dimensions, accounting for their interdependence.

5 A COORDINATED RESEARCH AGENDA

The reinforcing structure of the control deficit described in the preceding section implies that narrowly scoped interventions, pursued in isolation, will be insufficient—progress on any single dimension risks being undermined by stagnation in the others. Addressing the control deficit, therefore, requires a coordinated research agenda in which operational, technical, human, and governance lines of effort advance in parallel. Some dependencies impose sequencing constraints. Other efforts can and must proceed concurrently because the forcing function of adversary adoption does not permit waiting for earlier dependencies to be fully resolved. The research priorities presented below reflect this logic, addressing each dimension of the control deficit while accounting for its interactions.

5.1 Build Agents and Teams for Real-World Operations

The first strategic thrust focuses on developing operationally viable human-agent teaming paradigms that support cyber operations under real-world constraints. Moving from promising demonstrations to operational impact will require different strategies for offense and defense. Offensive operations can advance through human-AI teaming that compensates for current agent limitations, while defensive applications demand more fundamental technical progress before agents can be trusted in production environments.

5.1.1 Human-AI Teaming for Offensive Operations. The capability gaps documented in Section 4.1.1—multi-stage persistence, target identification, and consistency—are unlikely to be resolved by model capability alone in the near term. A more tractable path positions human-AI teaming as the operational paradigm for offensive cyber operations, with humans serving as mission orchestrators who compensate for current agent limitations while exploiting their strengths in speed, scale, and tactical execution. However, effective human-AI teaming in this context is not simply a matter of placing a human operator alongside an autonomous agent. Several open research problems must be addressed before this paradigm can be fielded effectively.

First, the distribution of control between the human and the agent across operational phases remains poorly understood. Current agents cannot independently sustain multi-stage operations, and the optimal points at which human direction is needed and the mechanisms by which agents should signal the need for intervention remain undefined. This is fundamentally a dynamic autonomy problem: the appropriate level of agent independence varies across phases of an operation and must adapt to operational conditions in real time (Hauptman, Mallick, et al. 2025).

Next, the identification-versus-execution gap creates a cognitive scaling challenge for human operators. If agents cannot reliably distinguish relevant information from useless noise, that function falls to the human orchestrator. At the scale and tempo enabled by agents, the volume of data this entails is likely to overwhelm human operators. Research is therefore needed on how to present agent observations and findings to operators in forms that support rapid triage and prioritization without overwhelming cognitive capacity.

Finally, oversight mechanisms must be designed to catch agent errors without negating the speed advantage that automation provides. If human verification is required for every action, the operational benefit of autonomy is lost; if verification is too infrequent, errors compound undetected. Identifying the appropriate level of oversight granularity and developing mechanisms that enable agents to flag low-confidence decisions for human review are open design problems with direct implications for operational effectiveness.

This agenda item will require collaboration among AI designers, human-computer interaction specialists, cognitive systems researchers, and operational cyber personnel to realize the integration of human/agentic cyber teams. Cyber practitioners are already integrating agents into daily workflows, making this research urgent and requiring immediate attention.

5.1.2 Bridging the Sim-to-Real Gap for Defensive Operations. In contrast to offensive operations, where human-AI teaming offers a near-term path to operational impact, the road to autonomous cyber defense is longer and requires more fundamental technical progress. In the near term, agents will

continue to provide efficiency improvements in process automation, querying and analysis, vulnerability discovery and patching, and training and preparedness for cybersecurity personnel (Forum 2024). Beyond this, three lines of research are necessary to close the gaps identified above and make operationally viable ACD agents a reality.

The sim-to-real gap demands a new generation of training environments that prioritize operational realism. Current platforms define observation and action spaces that diverge from those in production environments, preventing trained agents from transferring to real networks. Closing this gap requires environments that present agents with partial, noisy, and delayed visibility, characteristic of operational networks, rather than the complete state information available in current simulators. Equally important, observation spaces should be designed around the orchestration of existing defensive tools rather than forcing agents to re-learn the functionality those tools already provide (Oesch et al. 2024). Developing these environments will likely require data from actual network operations to calibrate fidelity, raising questions about data access, classification, and privacy that intersect with the governance challenges discussed earlier (Section 4.3).

The generalization constraints inherent in deep reinforcement learning represent a more fundamental research challenge. Fixed-size observation and action spaces prevent agents from transferring across network topologies or incorporating new defensive capabilities without architectural modification and costly retraining. Addressing this requires research into architectures that accommodate variable-size inputs and modular action spaces. Approaches such as graph-based observation representations (Li et al. 2026; King, Bowman, and Huang 2025) and hierarchical action frameworks allow agents to adapt to new environments without full retraining. These are active areas of research in the broader RL community, but their application to autonomous cyber defense remains largely unexplored.

Finally, the opacity of RL-based decision-making must be addressed before ACD agents can be trusted in operational settings. Post-hoc attribution methods can provide feature-level explanations for individual decisions, offering operators insight into which observations influenced a given action (Pakina, Sharma, and Pujari 2025). A more ambitious approach replaces opaque neural-network reward models with inherently interpretable ones, thereby making the reward structure itself transparent (Purves et al. 2024). LLMs provide an alternative approach to explainability by adding an explanation layer on top of RL-based policy networks or replacing these policy networks entirely (Chowdhry, Manero, and Sampalli 2025; Castro et al. 2025). Each approach entails trade-offs among computational cost, fidelity, and operational overhead, and none has yet been validated in operational defensive settings. Determining which methods provide explanations that operators find relevant, timely, and actionable is an open research question with direct implications for the trust on which adoption depends.

This agenda item will require collaboration among explainable AI (XAI), Game Theory, RL, and simulation design experts to create the realistic and fluid simulations needed to train and field effective cyber defense agents. As these agents are required in the near term, this research is ongoing and will remain a priority.

5.1.3 Reduce the Attack Surface. A new class of vulnerabilities is emerging in AI systems that operate with greater independence, and the challenge they pose is twofold. The frontier models that power agentic systems must themselves be hardened against manipulation, jailbreaking, and adversarial exploitation; a capable agent built atop a compromised model inherits and amplifies its weaknesses.

Beyond this, the supporting components that grant these models agency introduce a distinct class of system-level vulnerabilities. Both must be addressed to ensure that agents deliver value that exceeds the risk they introduce.

Frontier Model Security. An LLM's unification of commands and data is not a bug: it is an inherent property of how these systems operate, making jailbreaks and prompt injection resistant to architectural elimination. Despite this structural vulnerability, meaningful progress is being made towards hardening frontier models against manipulation. Defenses now operate at multiple levels. At the model level, safety post-training techniques align models to refuse harmful requests, while auxiliary classifiers monitor inputs and outputs to detect and block attempts that slip past initial safeguards (Cunningham et al. 2026). At the systems level, developers are adopting defense-in-depth architectures that layer real-time monitoring, account-level moderation, and rapid remediation pipelines to identify and patch universal jailbreaks as they emerge (Chen et al. 2024).

These investments are paying off: the UK AI Safety Institute (AISI) documented a 40-fold increase in expert effort required to jailbreak leading models for biological misuse queries over a six-month period (AISI 2025). Meanwhile, frontier models are setting new standards in robustness to prompt injections, achieving attack success rates (ASR) of just 1% in internal benchmarks (Anthropic 2025c). The trajectory is encouraging, but no deployed system has proven fully robust: every model AISI tested was eventually compromised, and a 1% ASR still represents meaningful risk. Continued research into both model-intrinsic robustness and architectural safeguards will be essential as these systems assume greater autonomy.

Agent Security. Securing agentic architectures requires attention to both the protocols that connect models to external tools and the design principles that govern agent deployment. At the infrastructure level, practices such as tool version pinning (Rhoads 2025) reduce the risk of tool injection attacks by ensuring that agents interact only with known, vetted tool implementations. Command injection vulnerabilities in MCP servers, meanwhile, are addressed through traditional software security best practices: input validation, least-privilege execution, and secure coding standards that predate the AI era but remain essential within it. Open-source developers are improving the Model Context Protocol itself, hardening it against misuse by adversaries and by well-meaning developers who might inadvertently introduce vulnerabilities (Kas 2025).

At the architectural level, frameworks such as the "Agents Rule of Two" offer practical guidance for limiting the "blast radius" when a compromise occurs (AI 2025). The framework, which recommends against agent deployments that provide simultaneous access to private data, exposure to untrusted content, and the ability to take consequential actions, offers a useful supplement to existing security best practices. Together, these measures reflect an emerging consensus: securing AI agents is not a single problem to be solved but a discipline to be practiced across multiple levels of abstraction in the cybersecurity stack.

This agenda item will require deep collaboration between industry and the frontier model providers responsible for the development of the most capable AI models.

5.2 Foster Trust in our New Agentic Partners

With the technical foundations addressed above, the next prerequisite for responsible adoption is trust that enables human operators to work effectively alongside these systems. Explainable AI (XAI) aims to build trust by demystifying the "black box," providing users with human-interpretable insight into how and why a model reaches its decisions (Dwivedi et al. 2023). Transparency alone, however, is not enough: explanations increase user trust only when operators perceive them as relevant to their own decisions (De Brito Duarte et al. 2023), and they are most effective when the system communicates not just its reasoning but also its confidence in that reasoning (Hauptman et al. 2024).

Despite these apparent benefits, numerous studies suggest that more does not necessarily mean better. Research on AI explanations in high-risk contexts shows that too much explainability can cognitively overload human operators (Hauptman, Mallick, et al. 2025), and work on the employment of AI agents for cybersecurity finds that in order for AI explanations to be well-received and effective, they need to occur using the same modalities that the rest of the team uses to communicate (Hauptman et al. 2024). Absent this, explanatory information is likely to be either missed or disregarded. Trust if fragile in another way as well: when an AI agent makes an incorrect decision, it erodes the trust that its human users have in its capabilities. Human-AI teaming research shows that traditional human-to-human trust repair strategies are far less effective in human-AI teaming; humans much more readily forgive human mistakes, often assuming that machines will make up for them (B. G. Schelble et al. 2024).

Given the above conclusions, future research must incorporate shared ethical ideologies into AI agents designed for cyber operations. Team trust and performance are heavily grounded in a team's ability to predict the behavior of all team members, based on shared values and beliefs (Flathmann et al. 2021). AI policymakers and designers need to work together to develop policies that reflect US values and interests, translate those policies into coded guardrails and responses, and provide explanations to human operators that communicate the ethical factors underlying an agent's decisions. When an AI agent makes a mistake, it must be designed not only to communicate what happened but also to perform tested trust-repair strategies with human operators to regain their confidence.

This agenda item will require collaboration among ethics, XAI, HCI, cognitive psychology, and behavioral science experts. Trust is a longer-term goal that requires a deeper understanding of the implications of human-agent collaboration.

5.3 Address the Policy Gap

Technical progress and trust mechanisms create the conditions for adoption, but adoption at scale requires governance frameworks that do not yet exist. Promoting norms, policy development, and international cooperation will be essential to reducing instability and managing the strategic implications of agentic AI in cyber operations. Section 4.3 highlighted key areas of law and policy that must be addressed to responsibly integrate AI into cyber operations.

These frameworks should be built from the bottom up. In the modern era of lightning-fast innovation and uncertain balances of power, international law has become increasingly dependent on the precedents of domestic law and policy (Slaughter and Burke-White 2007). Democratic nations must prioritize developing domestic governance frameworks for AI integration to shape their eventual

international implementation. As domestic standards for the acceptable design and deployment of agentic AI in cyber operations solidify, these standards can serve as the basis for international consensus.

To support the current administration's push for rapid agentic AI development, the government must establish a regulatory framework that defines the objectives, limits, and applicable law for AI in operational use. To date, the US approach to AI regulation has been largely decentralized and relies on voluntary compliance (Davtyan 2025). This allows for sector-specific regulations that account for each sector's needs and risk profile. An overly-comprehensive AI regulatory framework can be difficult to future-proof and hinder innovation (Davtyan 2025). However, there is a need to establish a baseline risk-based framework and standards that all sectors can utilize, particularly when developing AI to support operations. A good starting point is NIST's Artificial Intelligence Risk Management Framework, which outlines the critical steps and functions for organizations to assess and manage AI risks across design, development, and deployment (U.S. Department of Commerce 2023). Another good starting point is the European Union's Artificial Intelligence Regulation, a comprehensive legal framework for governing the development and deployment of AI systems, which includes the creation of new AI governance bodies (Cancela-Outeda 2024).

This agenda item will require collaboration among AI, national and international law, technology policy, and ethics experts. These policies have implications beyond national boundaries and will require both multidisciplinary and international collaboration. Policy should evolve as technology advances, and this is likely to remain an enduring requirement for years to come.

5.4 Manage Escalation Risks and Confront Dual-Use

The dual-use nature of agentic technology demands measures that operate at the boundary between legitimate and adversarial use. The same AI capabilities that enable legitimate cyber operations also empower malicious actors. Managing this dual-use dilemma requires both robust methods to assess when models cross dangerous capability thresholds and access frameworks that distinguish trusted operators from potential adversaries.

5.4.1 Create Better Evaluations to Mitigate Escalation Risks. To address concerns about the potential risks posed by AI advances in cybersecurity, most advanced model developers commit to a frontier AI safety framework — structured guidelines for anticipating and managing risks from their most capable models (Forum 2025). At their core, these frameworks define specific assessments and associated thresholds for dangerous cybersecurity capabilities (e.g., autonomous vulnerability discovery and exploit generation). They then prescribe various conditional commitments: if a model crosses a capability threshold, predefined mitigations are applied. Mitigations scale with capability, with the highest capability levels implying development halts under some frameworks. By committing in advance to the triggers for concern and the actions that follow, these frameworks aim to avoid ad hoc reasoning under commercial pressure when new dangerous capabilities emerge.

Frontier safety frameworks are limited by the degree to which their evaluations identify and measure relevant capabilities. Current assessments are multifaceted and increasingly robust, yet a gap is emerging between benchmark performance and real-world impact. Frontier model providers acknowledge that their current evaluation infrastructure is necessary but not sufficient for assessing

operational capability, noting that existing benchmarks lack active defenses, realistic noise, and the complexity that characterizes hardened targets (OpenAI 2025a). The consequences of this gap are already visible. Agents succeed in emulated environments while producing zero findings when tested against live enterprise networks (Lin et al. 2025; Singer et al. 2025a), demonstrating that strong benchmark performance can coexist with operational irrelevance. The reverse is also possible: constrained evaluation settings may understate the uplift that emerges when agents are embedded in realistic workflows, leaving current frameworks at risk of error in both directions.

Addressing these compounding evaluation failures requires a fundamental shift in approach. More difficult benchmarks alone are insufficient; the field needs evaluations that are difficult in the right ways, capturing the bottlenecks that currently constrain real-world attackers rather than abstract technical challenges that models may circumvent in practice. This means assessments that measure compound task completion across realistic attack chains, dynamic environments that adapt to model behavior rather than presenting static challenges, and baselines that establish what attackers can already accomplish without AI assistance. Where feasible, controlled studies of actual uplift provided to users with varying skill levels, and evaluations comparing agents with human cybersecurity professionals, offer promising methods for grounding results in operational reality (Lin et al. 2025). Without these investments, the frameworks designed to manage frontier risk will increasingly operate in the dark.

5.4.2 Deploy Differentiated Access Frameworks against Dual-Use. Model providers face an imperative to prevent their systems from facilitating malicious cyber operations, particularly as AI-assisted hacking lowers the barrier to sophisticated attacks. In response, most providers implement "guardrails"—post-training alignment procedures, reward modeling that discourages malicious use cases, and input/output classification systems designed to detect and block attempts to weaponize model capabilities. However, these protective measures create an operational dilemma for defensive actors and for legitimate offensive use cases. Red team operations require realistic offensive capabilities to stress-test organizational defenses, and vulnerability researchers need models that understand exploitation techniques to identify and remediate security flaws. Furthermore, military cyber operators will need unrestricted frontier models to conduct offensive operations that help preserve national security.

Removing safety guardrails through techniques such as abliteration is one potential remedy, but such approaches are limited to open-source models, which lag behind the frontier of capabilities. This tension, therefore, suggests the need for differentiated access frameworks for models. For national security use-cases, differentiated access might be achieved through partnerships with frontier labs to provide military- or government-specific model variants that omit civilian guardrails. The 2025 contract awards between the Chief Digital and Artificial Intelligence Office (CDAO) and the leading frontier labs suggest that such partnerships are possible (U.S. Department of Defense 2025), and efforts like OpenAI's Trusted Access Program (OpenAI 2026) and Anthropic's Project Glasswing (Anthropic 2026c) provide a blueprint for implementation, especially important in response to the potential large-scale vulnerabilities made public by Mythos (Anthropic 2026a). This program provides qualifying users with differentiated access to model capabilities relevant to cyber operations, allowing trusted actors to hold conversations and run agents that would otherwise be flagged as disqualifying malicious activity (OpenAI 2025d). This type of coordination represents the first steps toward consolidating

the gains from agentic AI for legitimate use cases while avoiding simultaneous augmentation of our adversaries' capabilities.

This agenda item lies at the intersection of cybersecurity, AI safety and evaluation, policy and governance, and international security, requiring coordinated approaches that address both the technical potential of these systems and the strategic risks they introduce. Future AI advances will immediately be evaluated by potentially nefarious actors for opportunities to scale, automate, or accelerate malicious activity, making the pace of defensive adaptation and governance a critical determinant of whether these systems stabilize or destabilize the cyber domain.

5.5 Coordinating the Agenda

The research priorities presented above are organized by dimension for clarity, but cannot be pursued in isolation without reproducing the very dynamics that define the control deficit. The reinforcing cycle identified in Section 4.5 propagates in two directions: forward, from technical foundations through trust to governance, and in return, from unmanaged dual-use risks through adversary adoption to renewed pressure on those same foundations. To be effective, this agenda organizes these priorities into two coordinated efforts designed to attack the self-reinforcing cycle from opposite directions. This coordinated approach is summarized in Figure 3.

Effort 1: Addressing the Forward Deficit Chain. The first effort addresses the "forward" direction of the cycle, where technical limitations create trust gaps that, in turn, stall governance. It contains two sequential dependencies that must be respected. The explainability research described in Sections 5.1.2 and 5.2 is a prerequisite for the trust mechanisms that underpin adoption. Operators cannot calibrate their trust in systems whose reasoning they cannot access, and trust repair strategies cannot be validated against systems that offer no window into why a failure occurred. Progress in explainable AI simultaneously addresses the technical opacity and trust dimensions, breaking the first link in the reinforcing chain. The second dependency runs from trust to governance: governance frameworks for autonomous cyber operations require operational experience to be well-calibrated, and that experience cannot accumulate at scale until organizations trust these systems enough to deploy them. The trust research described above, therefore, enables not only adoption but also the empirical grounding needed for governance to move beyond aspiration.

Effort 2: Managing the Return Pressure. The second effort confronts the return pressure that amplifies the cycle, where the absence of governance leaves dual-use risks unmanaged and compresses the timeline for all other work. Unlike the forward chain, the agenda items in this effort are structurally independent of each other and of the forward dependencies, which is precisely why they must be pursued in parallel with *Effort 1* rather than waiting for its foundations to mature. Differentiated access frameworks (Section 5.4.2) must be pursued now. The restrictions that constrain legitimate operators and adversaries alike cede advantage to actors who face no comparable constraints. Frontier capability evaluations (Section 5.4.1) must also advance in parallel because weakening benchmarks imply that the governance mechanisms designed to trigger mitigations at dangerous capability thresholds are losing their ability to detect them. Both of these efforts operate at the boundary between legitimate and adversarial uses and can proceed independently of the technical and trust foundations that are still under development.

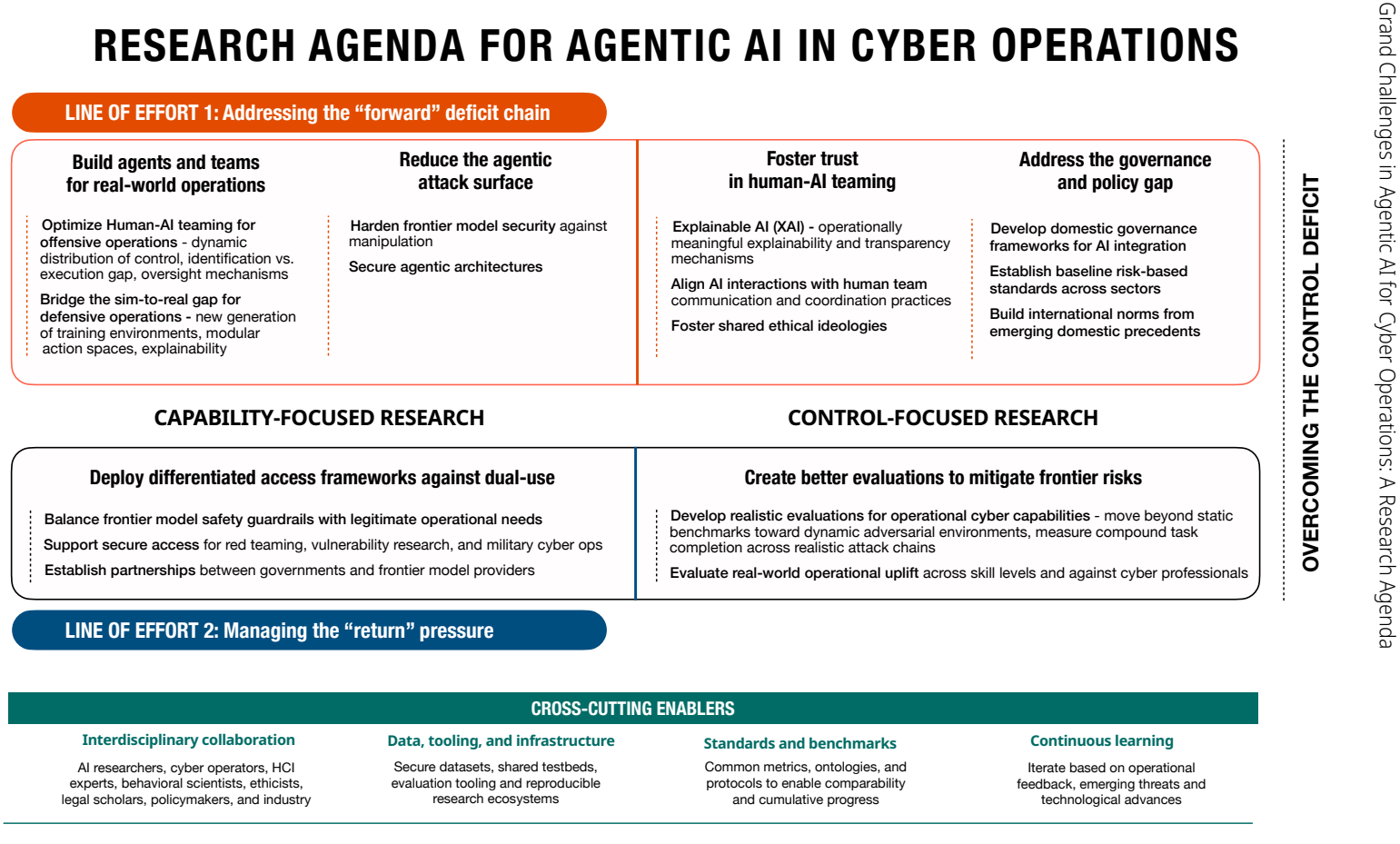

Figure 3. Research agenda for Agentic AI in cyber operations. The line of effort approach of the research agenda is depicted in this figure as having two parts, capability-focused research and control-focused research. The two efforts are intended to be conducted in parallel to overcome the control deficit.

The two efforts are not alternatives; they are interventions on the two directions of a single reinforcing cycle, and closing the control deficit requires addressing both.

Addressing the control deficit requires coordinated progress across multiple stakeholder communities, including AI researchers, cybersecurity practitioners, HCI researchers, military operators, frontier model developers, policy-makers, and governance institutions. Ultimately, the challenge is not only to integrate agentic AI into cyber operations, but to shape the conditions under which that integration occurs. The field must therefore move beyond isolated technical optimization toward coordinated socio-technical stewardship of increasingly autonomous cyber capabilities.

6 CONCLUSION

This paper argues that the rapid operationalization of agentic AI in cyberspace has produced a control deficit—a widening gap between what these systems are capable of and our ability to deploy them responsibly in operational cyber environments. We have characterized this deficit not as a collection of isolated challenges, but as a structured problem space spanning four interdependent dimensions: technical limitations, the expanded attack surface introduced by agentic architectures, a persistent trust deficit between human operators and autonomous systems, inadequate governance frameworks, and the escalation risks inherent to dual-use capabilities.

Critically, these dimensions do not exist independently. They form a reinforcing system of constraints and feedback loops that shape both the trajectory of adoption and the risks of misapplication. Technical opacity undermines operator trust; diminished trust slows integration into operational workflows; limited deployment constrains the empirical basis needed to inform governance; and governance gaps leave dual-use risks unmanaged amid persistent adversarial pressure. Taken together, these reinforcing dynamics ensure that localized or sequential solutions will fail to meaningfully close the control deficit.

Recognizing this structure is therefore not merely descriptive—it is foundational. It shifts the problem from filling discrete gaps to designing system-level interventions. The research agenda proposed here reflects this shift, prioritizing coordinated lines of effort that explicitly account for dependencies among technical, human, and institutional domains, rather than treating each dimension in isolation. The challenge ahead is thus not reducible to improvements in capability, trust, or governance alone, but lies in addressing the architecture of their interaction. Meeting this challenge will require sustained engagement across a broad set of multidisciplinary communities—spanning AI, cybersecurity, human–machine interaction, law, ethics, psychology, policy, and strategic studies, not in isolation or sequence, but as part of an integrated effort to reshape how agentic systems are developed, fielded, and governed.

Looking forward, the central question is not whether agentic systems will become embedded in cyber operations—they already are—but whether their integration will be guided by coherent frameworks that align technical behavior, human trust, and institutional control. In the absence of such alignment, the control deficit will continue to widen, with increasing consequences for operational stability and strategic risk. Addressing this gap ultimately requires a shift in perspective: from treating agentic AI as a capability to be optimized, to treating it as a socio-technical system to be governed under conditions of uncertainty and competition. Progress will depend on developing approaches that are

not only technically robust but also adaptively governable, cognitively aligned with human operators, and resilient under adversarial pressure. The work ahead, therefore, is not simply to improve these systems—but to ensure that, as they scale, our capacity to control, understand, and responsibly deploy them scales with them.

ABOUT THE AUTHORS

CPT Kyle Dotterrer is a Research Scientist at the Army Cyber Institute at West Point, NY. He holds a B.A. in Computer Science and Mathematics from Dartmouth College and an M.S. in Computational Data Science from Carnegie Mellon University. His current research interests include AI applications in cyber operations and AI-assisted software engineering.

MAJ Allyson Hauptman is a Senior Research Scientist at the Army Cyber Institute and an Assistant Professor at the United States Military Academy in the Department of Electrical Engineering and Computer Science. She holds a M.S. in Cyber Security from Tallinn Technical University and a PhD. in Human-Centered Computing from Clemson University, where her dissertation focused on the human factors that should influence dynamic autonomy levels for AI teammates. Her primary research areas include human-AI teaming, cybersecurity policy, and the application of international law to developing technologies.

CW3 Blane Richoux is a Research Scientist at the Army Cyber Institute at West Point, NY. He holds an M.P.S. in Information Security and Forensics (Homeland Security) from Pennsylvania State University and an M.S. in Cybersecurity from the Georgia Institute of Technology, whose capstone research focused on automating cybersecurity processes for resource-limited and edge environments. His current research interests include agentic AI applications for cyber defense, Human-AI teaming for cyberspace operations teams, security systems engineering, and human factors in cyber conflict.

LTC Matthew Corbett is a Senior Research Scientist and AI Research Team Leader at the Army Cyber Institute at West Point, NY. He holds an M.S. in Computer Science from the Georgia Institute of Technology and a Ph.D. in Computer Science from Virginia Tech. His research focuses on the application of artificial intelligence to command and control (C2) and mission command, including the use of eye-gaze and AI to assess and enhance tactical situational awareness. His work also examines the integration of AI into cybersecurity systems and the development of effective human–AI teaming in autonomous and cyber environments.

REFERENCES

Adkins, Heather, Gadi Evron, and Bruce Schneier. 2025. “Autonomous AI hacking and the future of cybersecurity,” October 8, 2025. https://www.csoonline.com/article/4069075/autonomous-ai-hacking-and-the-future-of-cybersecurity.html.

AI, Meta. 2025. “Agents Rule of Two: A Practical Approach to AI Agent Security.” Meta AI, October 31, 2025. https://ai.meta.com/blog/practical-ai-agent-security/.

AISI. 2025. “Frontier AI Trends Report by The AI Security Institute (AISI).” AI Security Institute. https://www.aisi.gov.uk/frontier-ai-trends-report.

Anthropic. 2024. “Introducing the Model Context Protocol,” November 25, 2024. https://www.anthropic.com/news/model-context-protocol.

Anthropic. 2025a. “Detecting and countering misuse of AI: August 2025,” August 27, 2025. https://www.anthropic.com/news/detecting-countering-misuse-aug-2025.

Anthropic. 2025b. “Disrupting the first reported AI-orchestrated cyber espionage campaign,” November 13, 2025. https://www.anthropic.com/news/disrupting-AI-espionage.

Anthropic. 2025c. “Mitigating the risk of prompt injections in browser use,” November 24, 2025. https://www.anthropic.com/research/prompt-injection-defenses.

Anthropic. 2025d. “Responsible Scaling Policy, Version 2.2,” May 14, 2025. https://www-cdn.anthropic.com/872c653b2d0501d6ab44cf87f43e1dc4853e4d37.pdf.

Anthropic. 2026a. “Assessing Claude Mythos Preview’s cybersecurity capabilities,” April 7, 2026. https://red.anthropic.com/2026/mythos-preview/.

Anthropic. 2026b. “Introducing Claude Opus 4.6,” February 5, 2026. https://www.anthropic.com/news/claude-opus-4-6.

Anthropic. 2026c. “Project Glasswing: Securing critical software for the AI era,” April 7, 2026. https://www.anthropic.com/glasswing.

Batool, Amna, Didar Zowghi, and Muneera Bano. 2025. “AI governance: a systematic literature review.” *AI and Ethics* 5 (3): 3265–3279. https://doi.org/10.1007/s43681-024-00653-w.

Biden, Joseph. 2025. *Executive Order on Strengthening and Promoting Innovation in the Nation’s Cybersecurity,* January 16, 2025. https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2025/01/16/executive-order-on-strengthening-and-promoting-innovation-in-the-nations-cybersecurity/.

Cancela-Outeda, Celso. 2024. “The EU’s AI act: A framework for collaborative governance.” *Internet of Things* 27 (October): 101291. https://doi.org/10.1016/j.iot.2024.101291.

Castro, Sebastián R., Roberto Campbell, Nancy Lau, Octavio Villalobos, Jiaqi Duan, and Alvaro A. Cardenas. 2025. “Large Language Models are Autonomous Cyber Defenders.” In *2025 IEEE Conference on Artificial Intelligence (CAI),* 1125–1132. https://doi.org/10.1109/CAI64502.2025.00195.

Chen, Sizhe, Julien Piet, Chawin Sitawarin, and David Wagner. 2024. “StruQ: Defending Against Prompt Injection with Structured Queries.” *arXiv* (September 25, 2024). https://doi.org/10.48550/arXiv.2402.06363.

Chowdhry, Hassan, Jaume Manero, and Srinivas Sampalli. 2025. “Evaluating AI Agents for Cyber Defense: A Comparison of Deep Reinforcement Learning and LLM Approaches.” In *Intelligent Data Engineering and Automated Learning – IDEAL 2025: 26th International Conference, Jaén, Spain, November 13–15, 2025, Proceedings, Part II,* 423–433. Berlin, Heidelberg: Springer-Verlag, November 13, 2025. https://doi.org/10.1007/978-3-032-10489-2_36.

Council, Swiss Federal. 1978. *Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I).* Geneva, Switzerland, July 12, 1978.

Cunningham, Hoagy, Jerry Wei, Zihan Wang, Andrew Persic, Alwin Peng, Jordan Abderrachid, Raj Agarwal, et al. 2026. “Constitutional Classifiers++: Efficient Production-Grade Defenses against Universal Jailbreaks.” *arXiv* (January 8, 2026). https://doi.org/10.48550/arXiv.2601.04603.

DARPA. 2016. “DARPA Celebrates Cyber Grand Challenge Winners.” DARPA News, August 5, 2016. Government. https://www.darpa.mil/news/2016/cyber-grand-challenge-winners.

DARPA. 2017. “DARPA Cyber Grand Challenge OS syscall library.” GitHub, February 1, 2017. https://github.com/CyberGrandChallenge/libcgc.

Davtyan, Tatevik. 2025. “The U.S. Approach to AI Regulation: Federal Laws, Policies, and Strategies Explained.” *Journal of Law, Technology, & the Internet* (Cleveland, OH) 16 (2): 223–259. https://scholarlycommons.law.case.edu/jolti/vol16/iss2/2/.

De Brito Duarte, Regina, Filipa Correia, Patrícia Arriaga, and Ana Paiva. 2023. “AI Trust: Can Explainable AI Enhance Warranted Trust?” *Human Behavior and Emerging Technologies* 2023:1–12. https://doi.org/10.1155/2023/4637678.

Duan, Wen, Christopher Flathmann, Nathan McNeese, Matthew J Scalia, Ruihao Zhang, Jamie Gorman, Guo Freeman, Shiwen Zhou, Allyson I. Hauptman, and Xiaoyun Yin. 2025. “Trusting Autonomous Teammates in Human-AI Teams - A Literature Review.” In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems,* 1–23. New York: ACM. https://doi.org/10.1145/3706598.3713527.

Dwivedi, Rudresh, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, et al. 2023. “Explainable AI (XAI): Core Ideas, Techniques, and Solutions.” *ACM Computing Surveys* 55 (9): 1–33. https://doi.org/10.1145/3561048.

Elvitigala, Don Samitha, Armağan Karahanoğlu, Andrii Matviienko, Laia Turmo Vidal, Dees Postma, Michael D Jones, Maria F. Montoya, et al. 2024. “Grand Challenges in SportsHCI.” In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems.* CHI ’24. Honolulu, HI, USA: Association for Computing Machinery. https://doi.org/10.1145/3613904.3642050.

Fang, Richard, Rohan Bindu, Akul Gupta, and Daniel Kang. 2024. “LLM Agents can Autonomously Exploit One-day Vulnerabilities.” *arXiv* (April 17, 2024). https://doi.org/10.48550/arXiv.2404.08144.

Fink, Gonen. 2025. “The Agentic AI Platform for the Agentic Workforce of the Future.” Palo Alto Networks Blog, October 28, 2025. https://www.paloaltonetworks.com/blog/2025/10/agentic-ai-platform-for-agentic-workforce-future/.

Flathmann, Christopher, Beau G. Schelble, Rui Zhang, and Nathan J. McNeese. 2021. “Modeling and Guiding the Creation of Ethical Human-AI Teams.” In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society,* 469–479. AIES ’21: AAAI/ACM Conference on AI, Ethics, and Society. Virtual Event USA: ACM, July 21, 2021. https://doi.org/10.1145/3461702.3462573.

Folkerts, Linus, Will Payne, Simon Inman, Philippos Giavridis, Joe Skinner, Sam Deverett, James Aung, et al. 2026. “Measuring AI Agents’ Progress on Multi-Step Cyber Attack Scenarios.” AI Security Institute, March 16, 2026. https://www.aisi.gov.uk/research/measuring-ai-agents-progress-on-multi-step-cyber-attack-scenarios.

Forum, Frontier Model. 2024. “Issue Brief: AI for Cyber Defense.” Frontier Model Forum, November 22, 2024. https://www.frontiermodelforum.org/updates/issue-brief-ai-for-cyber-defense/.

Forum, Frontier Model. 2025. *Frontier Capability Assessments.* Technical report. April 22, 2025. https://www.frontiermodelforum.org/technical-reports/frontier-capability-assessments/.

Friar, Joseph, and Phillip Payne. 2025. *Agentic Artificial Intelligence: Strategic Adoption in the U.S. Department of Defense.* Technical report. Cybersecurity and Information Systems Information Analysis Center (CSIAC), September 3, 2025. https://csiac.dtic.mil/technical-inquiries/notable/agentic-artificial-intelligence-strategic-adoption-in-the-u-s-department-of-defense/.

Gabarda, Florencio Cano. 2025. "Model Context Protocol (MCP): Understanding security risks and controls," July 1, 2025. https://www.redhat.com/en/blog/model-context-protocol-mcp-understanding-security-risks-and-controls.

Greshake, Kai, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. "Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection." *arXiv* (May 5, 2023). https://doi.org/10.48550/arXiv.2302.12173.

Hauptman, Allyson I., Rohit Mallick, Christopher Flathmann, and Nathan J. McNeese. 2025. "Human factors considerations for the context-aware design of adaptive autonomous teammates." *Ergonomics* 68 (4): 571–587. https://doi.org/10.1080/00140139.2024.2380341.

Hauptman, Allyson I., Beau Schelble, Christopher Flathmann, Rohit Mallick, and Nathan McNeese. 2025. "Ethical adaptation: exploring the use of adaptive autonomy in the design of ethical AI teammates in healthcare." *AI and Ethics* 5, no. 5 (October): 5397–5414. https://doi.org/10.1007/s43681-025-00782-w.

Hauptman, Allyson I., Beau G. Schelble, Wen Duan, Christopher Flathmann, and Nathan J. McNeese. 2024. "Understanding the influence of AI autonomy on AI explainability levels in human-AI teams using a mixed methods approach." *Cognition, Technology & Work* 26 (3): 435–455. https://doi.org/10.1007/s10111-024-00765-7.

Hicks, Kathleen. 2023. *DoD Directive 3000.09 Autonomy in Weapon Systems.* Washington, DC, January 25, 2023. https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

Joint Chiefs of Staff. 2022. *Joint Publication 3-12, Cyberspace Operations.* Washington, DC. https://jdeis.js.mil/jdeis/index.jsp?pindex=27&pubId=791.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling Laws for Neural Language Models." *arXiv* (January 23, 2020). https://doi.org/10.48550/arXiv.2001.08361.

Kas, Pieter. 2025. "SEP-1932: DPoP Profile for MCP #1932," December 5, 2025. https://github.com/modelcontextprotocol/modelcontextprotocol/pull/1932.

King, Isaiah J., Benjamin Bowman, and H. Howie Huang. 2025. "Automated Cyber Defense with Generalizable Graph-based Reinforcement Learning Agents." *arXiv* (September 19, 2025). https://doi.org/10.48550/arXiv.2509.16151.

Li, Yu, Sizhe Tang, Rongqian Chen, Fei Xu Yu, Guangyu Jiang, Mahdi Imani, Nathaniel D. Bastian, and Tian Lan. 2026. "ACDZero: Graph-Embedding-Based Tree Search for Mastering Automated Cyber Defense." *arXiv* (January 5, 2026). https://doi.org/10.48550/arXiv.2601.02196.

Li, Ziyang, Saikat Dutta, and Mayur Naik. 2025. "IRIS: LLM-Assisted Static Analysis for Detecting Security Vulnerabilities." *arXiv* (April 6, 2025). https://doi.org/10.48550/arXiv.2405.17238.

Lin, Justin W., Eliot Krzysztof Jones, Donovan Julian Jasper, Ethan Jun-shen Ho, Anna Wu, Arnold Tianyi Yang, Neil Perry, et al. 2025. "Comparing AI Agents to Cybersecurity Professionals in Real-World Penetration Testing." *arXiv* (December 10, 2025). https://doi.org/10.48550/arXiv.2512.09882.

Lutnick, Howard. 2025. "Statement from U.S. Secretary of Commerce Howard Lutnick on Transforming the U.S. AI Safety Institute into the Pro-Innovation, Pro-Science U.S. Center for AI Standards and Innovation | U.S. Department of Commerce," June 3, 2025. https://www.commerce.gov/news/press-releases/2025/06/statement-us-secretary-commerce-howard-lutnick-transforming-us-ai.

Manning, Courtney. 2025. "Cloud of War: The AI Cyber Threat to U.S. Critical Infrastructure." American Security Project, October 9, 2025. https://www.americansecurityproject.org/cloud-of-war-the-ai-cyber-threat-to-u-s-critical-infrastructure/.

McNeese, Nathan J., Mustafa Demir, Erin K. Chiou, and Nancy J. Cooke. 2021. "Trust and Team Performance in Human–Autonomy Teaming." *International Journal of Electronic Commerce* 25 (1): 51–72. https://doi.org/10.1080/10864415.2021.1846854.

Mirishli, Shahmar. 2024. "Ethical Implications of AI in Data Collection: Balancing Innovation with Privacy." *ANCIENT LAND* 6, no. 8 (August 30, 2024): 40–55. https://doi.org/10.36719/2706-6185/38/40-55.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, et al. 2015. "Human-level control through deep reinforcement learning." *Nature* 518, no. 7540 (February): 529–533. https://doi.org/10.1038/nature14236.

Nasr, Milad, Nicholas Carlini, Chawin Sitawarin, Sander V. Schulhoff, Jamie Hayes, Michael Ilie, Juliette Pluto, et al. 2025. "The Attacker Moves Second: Stronger Adaptive Attacks Bypass Defenses Against Llm Jailbreaks and Prompt Injections." *arXiv* (October 10, 2025). https://doi.org/10.48550/arXiv.2510.09023.

Oesch, Sean, Phillipe Austria, Amul Chaulagain, Brian Weber, Cory Watson, Matthew Dixson, and Amir Sadovnik. 2024. "The Path To Autonomous Cyber Defense." *arXiv* (April 12, 2024). https://doi.org/10.48550/arXiv.2404.10788.

OpenAI. 2023. “OpenAI’s Approach to Frontier Risk: An Update for the UK AI Safety Summit,” October 26, 2023. https://openai.com/global-affairs/our-approach-to-frontier-risk/.
OpenAI. 2025a. “Addendum to GPT-5.2 System Card: GPT-5.2-Codex.” OpenAI Deployment Safety Hub, December 18, 2025. https://deploymentsafety.openai.com/gpt-5-2-codex.
OpenAI. 2025b. “Introducing Aardvark: OpenAI’s agentic security researcher.” Openai.com, November 20, 2025. https://openai.com/index/introducing-aardvark/.
OpenAI. 2025c. “Preparedness Framework V2” (April 15, 2025). https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf.
OpenAI. 2025d. “Strengthening cyber resilience as AI capabilities advance,” December 22, 2025. https://openai.com/index/strengthening-cyber-resilience/.
OpenAI. 2026. “Introducing Trusted Access for Cyber,” March 23, 2026. https://openai.com/index/trusted-access-for-cyber/.
Pakina, Anil Kumar, Ashwin Sharma, and Mangesh Pujari. 2025. “AI Governance via Explainable Reinforcement Learning (XRL) for Adaptive Cyber Deception in Zero-Trust Networks.” *Journal of Information Systems Engineering and Management* 10 (43): 98–115. https://doi.org/10.52783/jisem.v10i43s.8308.
Purves, Tom, Konstantinos G. Kyriakopoulos, Siân Jenkins, Iain Phillips, and Tim Dudman. 2024. “Causally aware reinforcement learning agents for autonomous cyber defence.” *Knowledge-Based Systems* 304:112521. https://doi.org/10.1016/j.knosys.2024.112521.
Qi, Xiangyu, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. 2023. “Visual Adversarial Examples Jailbreak Aligned Large Language Models.” *arXiv* (August 16, 2023). https://doi.org/10.48550/arXiv.2306.13213.
Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. “Language Models are Unsupervised Multitask Learners,” https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
Rehberger, Johann. 2023. “Hacking Google Bard - From Prompt Injection to Data Exfiltration.” Embrace The Red, November 3, 2023. https://embracethered.com/blog/posts/2023/google-bard-data-exfiltration/.
Rhoads, Thomas. 2025. “SEP-1766: Digest-Pinned Tool Versioning and Interceptor-Based Validation in MCP.” GitHub, November 5, 2025. https://github.com/modelcontextprotocol/modelcontextprotocol/issues/1766.
Russel, Stuart, and Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach.* 4th ed. Pearson.
Şahin, Emrullah, Naciye Nur Arslan, and Durmuş Özdemir. 2025. “Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning.” *Neural Computing and Applications* 37 (2): 859–965. https://doi.org/10.1007/s00521-024-10437-2.
Schelble, Beau, Caitlin Lancaster, Wen Duan, Rohit Mallick, Nathan Mcneese, and Jeremy Lopez. 2023. “The Effect of AI Teammate Ethicality on Trust Outcomes and Individual Performance in Human-AI Teams.” In *Proceedings of the 56th Hawaii International Conference on System.* Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2023.040.
Schelble, Beau G., Jeremy Lopez, Claire Textor, Rui Zhang, Nathan J. McNeese, Richard Pak, and Guo Freeman. 2024. “Towards Ethical AI: Empirically Investigating Dimensions of AI Ethics, Trust Repair, and Performance in Human-AI Teaming.” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 66, no. 4 (April): 1037–1055. https://doi.org/10.1177/00187208221116952.
Schmitt, Michael N., ed. 2013. *Tallinn Manual on the International Law Applicable to Cyber Warfare.* 1st ed. Cambridge University Press, March 7, 2013. https://doi.org/10.1017/CBO9781139169288.
Shellphish. 2016. “Mechanical Phish: @shellphish’s Cyber Reasoning System for the DARPA Cyber Grand Challenge.” GitHub. https://github.com/mechaphish.
Shoshitaishvili, Yan, Antonio Bianchi, Kevin Borgolte, Amat Cama, Jacopo Corbetta, Francesco Disperati, Audrey Dutcher, et al. 2018. “Mechanical Phish: Resilient Autonomous Hacking.” *IEEE Security & Privacy* 16, no. 2 (April): 12–22. https://doi.org/10.1109/MSP.2018.1870858.
Singer, Brian, Keane Lucas, Lakshmi Adiga, Meghna Jain, Lujo Bauer, and Vyas Sekar. 2025a. “Incalmo: An Autonomous LLM-assisted System for Red Teaming Multi-Host Networks.” *arXiv* (November 22, 2025). https://doi.org/10.48550/arXiv.2501.16466.
Singer, Brian, Keane Lucas, Lakshmi Adiga, Meghna Jain, Lujo Bauer, and Vyas Sekar. 2025b. “On the Feasibility of Using LLMs to Execute Multistage Network Attacks.” Version: 2, *arXiv* (March 6, 2025). https://doi.org/10.48550/arXiv.2501.16466.
Slaughter, Anne-Marie, and William Burke-White. 2007. “The Future of International Law is Domestic (or, The European Way of Law).” In *New Perspectives on the Divide Between National and International Law,* edited by Janne E. Nijman and André Nollkaemper. Oxford University Press, November. https://doi.org/10.1093/acprof:oso/9780199231942.003.0006.
Song, Jia, and Jim Alves-Foss. 2016. “The DARPA Cyber Grand Challenge: A Competitor’s Perspective, Part 2.” *IEEE Security & Privacy* 14, no. 1 (January): 76–81. https://doi.org/10.1109/MSP.2016.14.
Stackpole, Beth. 2026. “Agentic AI, explained | MIT Sloan,” February 18, 2026. https://mitsloan.mit.edu/ideas-made-to-matter/agentic-ai-explained.

Standen, Maxwell, Martin Lucas, David Bowman, Toby Richer, Junae Kim, and Damian Marriott. 2021. "CybORG: A Gym for the Development of Autonomous Cyber Agents." *arXiv* (August 20, 2021). https://doi.org/10.48550/arXiv.2108.09118.

Stephens, Nick, John Grosen, Christopher Salls, Andrew Dutcher, Ruoyu Wang, Jacopo Corbetta, Yan Shoshitaishvili, Christopher Kruegel, and Giovanni Vigna. 2016. "Driller: Augmenting Fuzzing Through Selective Symbolic Execution." In *Proceedings 2016 Network and Distributed System Security Symposium.* Network and Distributed System Security Symposium. San Diego, CA: Internet Society. https://doi.org/10.14722/ndss.2016.23368.

Stuckey, Dane. 2025. "[Post discussing prompt injection risks in ChatGPT agent and web browsing agents]." X. Posted by @cryps1s, October 22, 2025. https://x.com/cryps1s/status/1981037851279278414.

Trump, Donald J. 2025. *Removing Barriers to American Leadership in Artificial Intelligence,* January 23, 2025. https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/.

U.S. Department of Commerce. 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* V. 1.0. Washington, D.C., January. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

U.S. Department of Defense. 2025. "CDAO Announces Partnerships with Frontier AI Companies to Address National Security Mission." Chief Digital and Artificial Intelligence Office, July 14, 2025. https://www.ai.mil/Latest/News-Press/PR-View/Article/4242822/cdao-announces-partnerships-with-frontier-ai-companies-to-address-national-secu/.

Vyas, Sanyam, John Hannay, Andrew Bolton, and Pete Burnap Burnap. 2023. "Automated Cyber Defence: A Review." *arXiv* (March 8, 2023). https://doi.org/10.48550/arXiv.2303.04926.

Wei, Alexander, Nika Haghtalab, and Jacob Steinhardt. 2023. "Jailbroken: How Does LLM Safety Training Fail?" *arXiv* (July 5, 2023). https://doi.org/10.48550/arXiv.2307.02483.

Zhu, Yuxuan, Antony Kellermann, Akul Gupta, Philip Li, Richard Fang, Rohan Bindu, and Daniel Kang. 2025. "Teams of LLM Agents can Exploit Zero-Day Vulnerabilities." *arXiv* (March 30, 2025). https://doi.org/10.48550/arXiv.2406.01637.

Zou, Andy, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. "Universal and Transferable Adversarial Attacks on Aligned Language Models." *arXiv* (December 20, 2023). https://doi.org/10.48550/arXiv.2307.15043.

Received 23 January 2026; Revised 8 April 2026; Accepted 4 May 2026