SPECIAL COMPETITIVE STUDIES PROJECT

NOVEMBER 2026

Changing the Game

Assessing Agentic AI's Impact on Cyber Competitions and Cyber Operations

JAMES RYSEFF AND KACI MCBRAYER

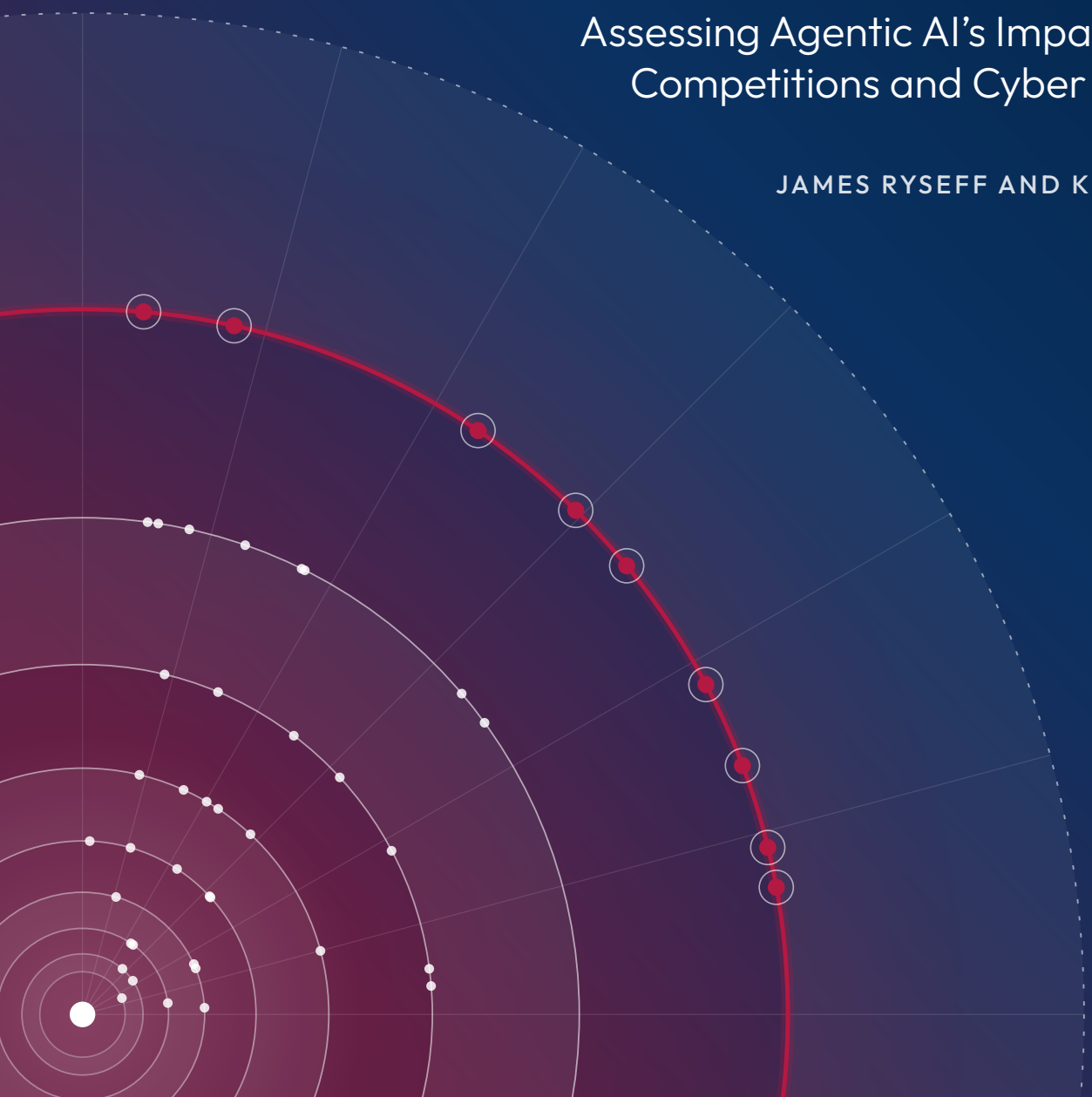


Table of Contents

Executive Summary 3

Scope and Method 5

Part 1: Why Agentic AI Is a Step-Change in Offensive Cyber 6

Part 2: A Framework for Cyber Operations, Reshaped by Human-Machine Teaming 9

Part 3: What Agentic AI Is Good At, and Why 11

Part 4: What Agentic AI Is Bad at and Why 14

Part 5: Takeaways 18

Part 6: Recommendations 21

Authors 23

Acknowledgements 23

Executive Summary

A Step-Change in Offensive Capabilities

Agentic AI represents a critical step-change in offensive cyber operations. The cyber threat landscape is no longer defined by models that merely answer questions; it is now increasingly dominated by autonomous agents that aggressively pursue increasingly complex objectives. Evidence from elite cyber competitions—environments that compress and expose the technical demands of cyber conflict—reveals a drastic shift in participant capabilities. Challenges that once stymied advanced tooling are now solved outright in a matter of hours or even minutes. Consequently, these autonomous capabilities have rendered some formats for traditional cyber competitions effectively obsolete, as agents can operate without human involvement entirely to secure the objective. We are rapidly approaching an era of machine-speed warfare, and our defensive posture must pivot accordingly.

Operational Strengths and Current Limitations

Currently, agentic AI operates with incredible speed and effectiveness when a problem is bounded and verifiable. It excels in white-box bug-finding, rapid exploit generation, large-scale data forensics, and obscure-technique research. Models are now capable of surfacing dozens of high-severity vulnerabilities across millions of lines of code within hours. However, the technology still contains important weaknesses. Most notably, agentic AI lacks stealth; interviewees overwhelmingly described it as "loud" and it routinely triggers intrusion detection systems during lateral movement. It also sometimes struggles with genuinely novel problems that demand creative reasoning, frequently hallucinating or reward hacking when it cannot find a clear path forward. Furthermore, agents lack the tacit mental models that experienced human operators use to intuitively navigate unfamiliar networks and often falters when executing long, multi-step exploits that require sustained coherence and judgment.

Takeaways

Agentic AI is actively reshaping the cyber kill chain by compressing the time and cost required for offensive operations. Its impact is most profound at the beginning and end of operations—specifically during the Initial Access/Foothold and Mission Execution phases—where bounded vulnerability discovery and rapid weaponization are paramount. While agents are currently less effective during the Reconnaissance/Staging

and Movement/Escalation phases where operational stealth is frequently required, they remain a massive force multiplier overall. By absorbing the technical implementation and "grunt work," agents free elite human operators to focus on overarching strategy, orchestration, and oversight. This human-machine symbiosis is transforming the productivity of cyber operations, enabling a single human operator to direct a swarm of agents to accomplish tasks that once required an entire team of specialists.

Recommendations

Preserving American cyber dominance will require swift leadership across three critical fronts. First, the federal government must redouble its efforts to ensure that national security operators have unhindered access to the most advanced models, backed by robust compute infrastructure across all classification levels. The United States simply cannot afford to be technologically outgunned by adversaries when executing its most important and most sensitive national security missions. Second, the government should orchestrate a massive, Y2K-style initiative to incentivize and coordinate private-sector preparedness, ensuring vulnerable organizations can leverage a broader ecosystem of automated, scalable defense. Finally, the United States must reinvent its cybersecurity competitions and training pipelines. Traditional Capture-the-Flag formats are rapidly losing relevance in an agentic era, eliminating a historic onboarding ramp that discovered promising new talent and taught them the fundamental tools of their trade. To attract the next generation of cybersecurity professionals we must fund and design new environments centered around human-AI teaming, teaching the next generation of defenders how to command, supervise, and outmaneuver autonomous swarms.

Scope and Method

This paper draws on fourteen interviews conducted in July and August 2026 with practitioners across cyber competitions and operations, together with a short, written survey of five additional cyber competitors. Consistent with the confidentiality terms of the engagements, subjects are described only by function and are not attributed by name or affiliation. We draw primarily on participants in cyber competitions because those competitions compress many of the technical and operational demands of cyber conflict into an environment where agentic capabilities can be directly observed and compared. Participants in these competitions are also often able to speak more freely compared to offensive and defensive cyber practitioners working for governments, criminal organizations, or corporations.

Part 1: Why Agentic AI Is a Step-Change in Offensive Cyber

AI is changing the unit of cyber power from the individual operator to the autonomous agent. It is tempting to read recent progress simply as a better model replacing a worse one, but that misses the point. Models have evolved from a system that answers your questions to one that goes off and pursues an objective on its own, with barely any supervision—you hand a challenge to an agent, walk away, and come back to a finished solution.

A few examples convey the scale. One competitor described spinning up as many as twenty concurrent agent instances to solve a particularly difficult CTF challenge, waiting an hour, and returning to an agent having identified an effective solution. Another built a tool that races roughly thirty models at a challenge at once and takes whichever finishes first. That approach was echoed almost verbatim in the survey data, where one respondent credited a custom orchestrator with "mass parallelization allowing for different approaches and models to work on the same challenges." Finally, an expert tool-builder reported that tooling which used to choke on certain challenges can now solve them outright given a two-day window. The net impact of this has been to render some types of cyber competitions effectively obsolete, since agents can solve all but the most difficult challenges without human involvement.[1]

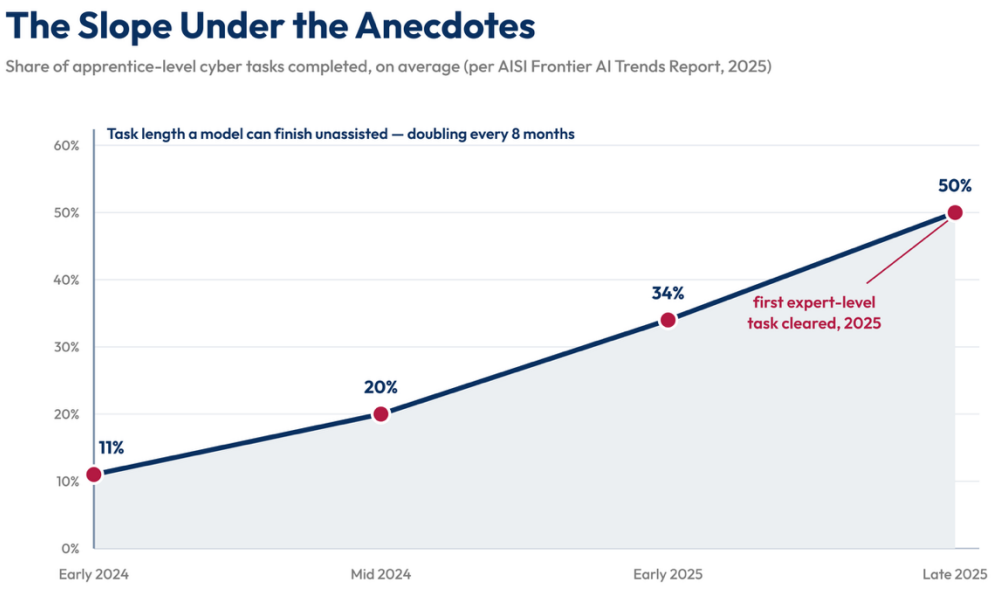
External reporting points in the same direction. In November 2025, Anthropic reported the first publicly documented case of an AI system autonomously conducting multi-step attacks against well-defended targets in the wild. Anthropic's account described attackers using AI's agentic capabilities not merely as an advisor but to execute the operation itself—by Anthropic's estimate the AI performed 80–90% of the tactical work for the campaign, with humans intervening at only four to six critical decision points per target, and at peak the system issued thousands of requests, often several per second.[2] Although the attack was notable less for any novel technique than for the autonomy of its orchestration, it nonetheless marked a threshold in the evolution of cybersecurity threats.[3] This incident was followed by a July 2026 incident where Hugging Face disclosed (and OpenAI later confirmed) an intrusion into its production infrastructure that it

[1] Announcing the Save CTFs Fund, Michael Debono (2026).
[2] Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign, Anthropic (2025).
[3] Christopher Covino, The Emergence of Autonomous Cyber Attacks: Analysis and Implications, Institute for AI Policy and Strategy (2025).

described as driven end to end by an autonomous AI agent system, running many thousands of actions across a swarm of short-lived sandboxes—what the company called a match for the "agentic attacker" scenario the industry had been forecasting.[4]

Why does cyber lend itself to this more than most domains? Because tasks in cyber competitions usually have a checkable answer—a flag captured, access to the command shell obtained, an exploit that works. That verifiability is the whole engine: a checkable outcome is exactly what you need to reward a model in training, so competition-style problems convert almost directly into training fuel. Organizations such as OpenAI,[5] Anthropic,[6] and The UK AI Security Institute (AISI)[7] have leveraged capture-the-flag exercises to train new models or evaluate the proficiency of advanced AI models in cybersecurity for exactly this reason.

The AISI data also puts a slope on the curve of how quickly these agents have improved. AISI found that models could complete apprentice-level cyber tasks about 50% of the time on average by late 2025, up from just over 10% in early 2024. Additionally, in 2025 it tested the first model able to complete expert-level tasks that would normally demand more than 10 years of human experience. By its most recent estimate, the length of cyber task a model can finish unassisted is roughly doubling every eight months.[8]



[4] Security Incident Disclosure — July 2026, Hugging Face (2026).
[5] Gpt-oss-120b & gpt-oss-20b Model Card, OpenAI (2025).
[6] Claude is competitive with humans in (some) cyber competitions, Anthropic (2025).
[7] Frontier AI Trends Report, AI Security Institute (2025).
[8] Frontier AI Trends Report, AI Security Institute (2025).

Figure 1. Source: AISI Frontier AI Trends Report (2025)[9]

All of which yields the paper's central claim. As of mid-2026, agentic AI has reached an uneven but definable frontier in cyber skill. It has effectively saturated one class of task—the bounded, verifiable kind—while remaining more limited on the novel, multi-step, discretion-heavy black-box kind. Consequently, agentic AI is poised to transform the productivity of the cybersecurity workforce - both attackers and defenders - and to transform how these professionals are taught their craft.

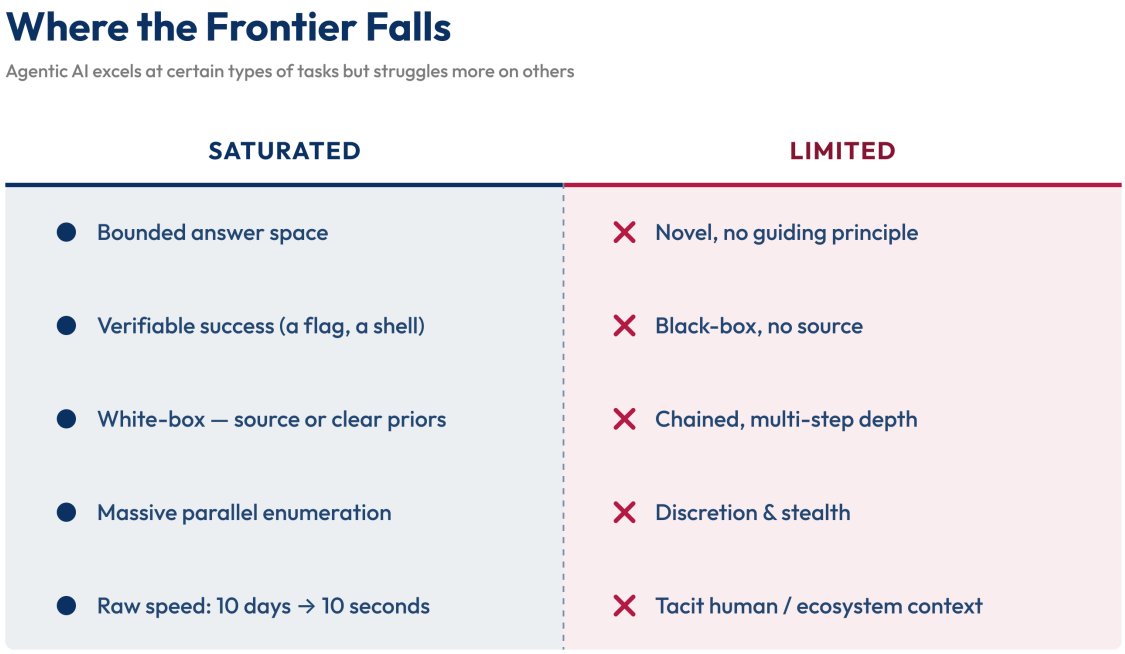


Figure 2. Source: SCSP Analysis

[9] Frontier AI Trends Report, AI Security Institute (2025).

Part 2: A Framework for Cyber Operations, Reshaped by Human-Machine Teaming

To organize where these tools work and where they struggle, we collapse the MITRE ATT&CK[10] framework into four phases: Reconnaissance and Staging, Initial Access and Foothold, Movement and Escalation, and Mission Execution.

These four phases capture the basic progression of a cyber operation from identifying a target to achieving the attacker's ultimate objective. Reconnaissance and Staging covers the work that occurs before an intrusion begins, including gathering information about a target's people, systems, and exposed infrastructure; identifying potential vulnerabilities; acquiring domains, credentials, malware, or other tooling; and configuring the infrastructure needed to support an attack. Initial Access begins when the attacker attempts to enter the target environment, whether through phishing, stolen credentials, exploitation of a vulnerability, or another initial-access technique, and then establishes a reliable presence on a compromised system. Movement and Escalation encompasses the steps used to expand that presence: discovering internal systems and accounts, obtaining additional privileges, moving laterally to other machines or networks, evading defenses, and creating mechanisms for persistent access. Finally, Mission Execution is the phase in which the attacker carries out the operation's underlying purpose, such as locating and collecting sensitive information, exfiltrating data, disrupting or destroying systems, deploying ransomware, manipulating data or services, or otherwise producing the intended operational effect.

[10] MITRE Attack Framework, MITRE (last accessed 2026).

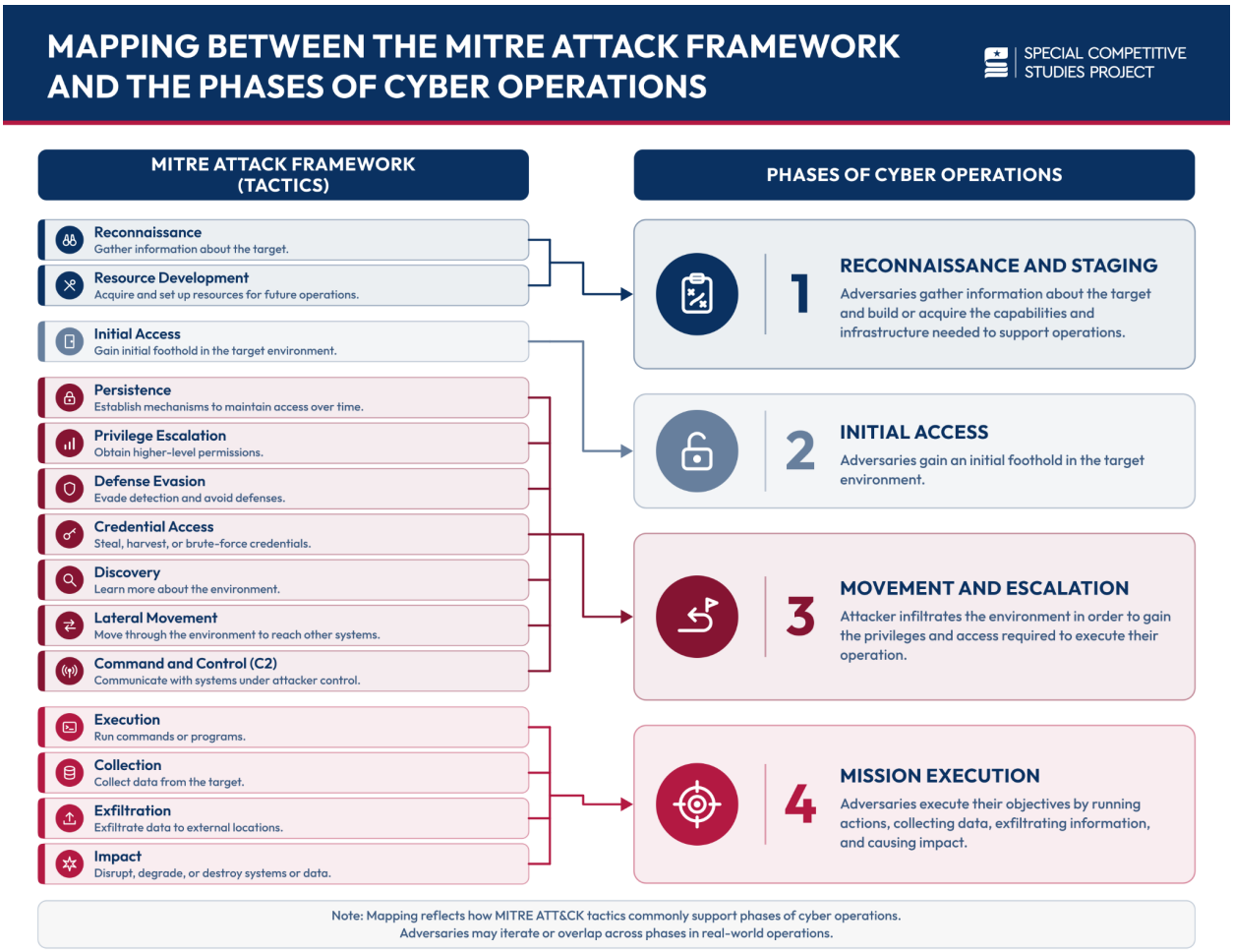


Figure 3. Source: SCSP Analysis

Adopting this framework illustrates how agentic AI is redrawing the cyber kill chain by shifting the division of labor between humans and machines. The result is not simply the old kill chain running faster; it is a fundamentally restructured one.

Part 3: What Agentic AI Is Good At, and Why

Across the interviews and the survey ratings, one theme repeatedly emerges: agentic AI is at its best when a problem is bounded (the answer space is constrained), verifiable (success can be clearly evaluated as a binary yes/no), and there is source code or a clear prior to work from.

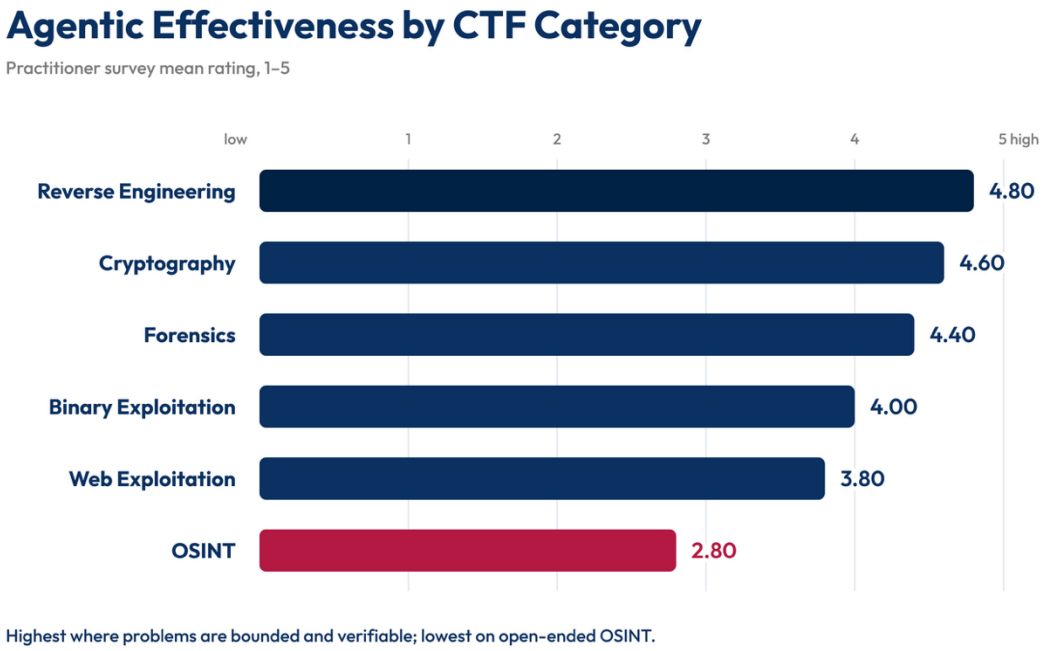


Figure 4. Source: SCSP Survey Data

White-box vulnerability discovery

The most striking single result came from an interviewee who specialized in tool building. His bug-bounty system, when pointed at a roughly five-million-line Ruby on Rails codebase, surfaced seventeen high-severity findings in about two hours, each validated with a working proof-of-concept. Other interviewees confirmed similar results under these optimal conditions — access to the full source codebase, a checkable success condition, and a bounded target—and it is precisely the class of white-box vulnerability the Hugging Face attackers exploited at machine speed.

Exploit generation

Interview subjects and survey data converged here finding that agentic AI had very high effectiveness ratings on exploitation-adjacent phases. A challenge author described handing a model a roughly thirty-megabyte executable for an old game and watching it locate a stack buffer overflow in about ten seconds and write the payload—work he estimated would have taken a person a week. A competitor described the scripted version: give the agent an nmap scan, a known vulnerability, and the network layout, and it reliably produces a working payload.

Large-scale forensics and data processing

Several interviewees rated agents highly on artifact-heavy work—packet-capture triage, steganography, forensics—where the job is to pick anomalous signal out of a very large haystack. The survey data mirrored this finding, with forensics among the higher-rated categories. The Hugging Face disclosure is a live demonstration of the defensive version of the same strength: to reconstruct what a swarm of tens of thousands of automated actions had done, the company ran LLM-driven analysis agents over an attacker log of more than 17,000 events and, by its own account, did in hours what would normally take days.[11]

Obscure-technique research at speed and scale

Several interviewees noted agentic AI's impressive ability to solve some of the most difficult CTF challenges for humans - ones that hinge on recognizing an unusual vulnerability or exploitation technique that even experienced practitioners may rarely encounter. Agents can research unfamiliar methods quickly, match them to the problem at hand, enumerate possibilities exhaustively, and operate far faster than a human. Speed and scale are the throughline: work that would take a person ten hours or ten days can collapse to something like ten seconds. Study participants summarized the strength as "the ability to operate at a scale never before seen" and the claim that the tools "can analyze better than any human now." External benchmarking agrees; in one evaluation, agents solved each CTF challenge in under an hour with little guidance, at a total cost under US$20 across all of them.[12] A commercial red-team agent pushed the same combination of speed and cost further, reportedly placing in the top 1% across six

[11] Security Incident Disclosure — July 2026, Hugging Face (2026).
[12] Benjamin Sperisen, et al., AI Agents Put Offensive Cyber Within Reach of Novices, RAND Corporation at 1 (2026).

competition platforms—ahead of more than 125,000 human competitors—at an average cost of roughly US$13 and under two hours per event.[13]

Put together, these clusters describe a tool that has more or less saturated the bounded-and-verifiable portion of the offensive workflow. The limits show up exactly where those properties no longer apply.

[13] Pavel Gurvich, Inside the Top 1%: Engineering Tenzai's AI Hacker to Compete With Elite Humans, Tenzai (2026).

Part 4: What Agentic AI Is Bad at and Why

The preceding sections show that agentic AI is already capable of materially accelerating a range of cyber tasks - particularly when the objective is clear, the environment is bounded, and success can be readily verified. But those strengths are only part of the picture. Our interviews also revealed some conditions under which agentic AI proves less effective. Understanding these weaknesses is essential to assessing where agentic AI is likely to have the greatest near-term impact—and where human expertise and judgement remain difficult to replace.

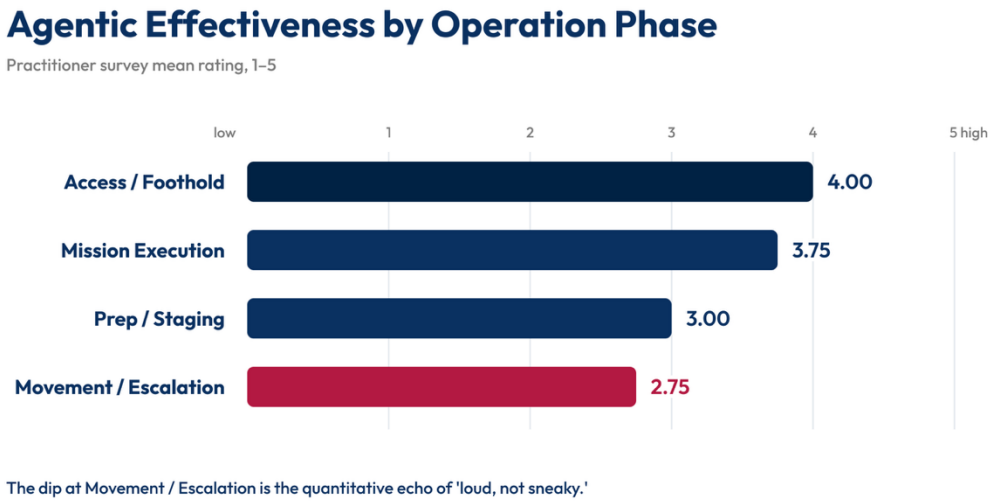


Figure 5. Source: SCSP Survey Data

Stealth — 'loud, not sneaky'

The most robust weakness in the whole corpus is stealth. Many of our interviewees on cyber competition teams described agents as “loud” and “not sneaky”. This took several forms - they take actions which intrusion detection is built to catch, leave their artifacts lying around, and poke at things they should leave alone. As one interviewee put it “I pitched the idea for a CTF challenge where the goal is to be undetected - that is the worst thing for an AI to do.” Another commented “It gets caught or noticed more often … even if you tell it to be sneaky, it doesn’t have the knowledge to know how.”

These characteristics led our interviewees to express reluctance to use agentic AI for actions in the lateral movement stage of cyber operations - at least without significant

human oversight and control. Even if the models were to improve, they worried that agentic AI could still represent a threat to their operation at this stage. As one put it, "in an actual operation, I wouldn't trust the AI not to blow our cover and get us caught. AI could understand how to be quiet, but it doesn't always do what you say. How can I trust it?" This is not to diminish the utility of agentic AI even within the lateral movement stage - under appropriate human control and supervision, agents can meaningfully assist human operators. Overall, our survey respondents rated agentic AI's utility in the "Lateral Movement/Privilege Escalation" stage at 2.75 on a 5 point scale. While this was the lowest of the four stages in cyber operations, it still demonstrates that agentic AI is having a meaningful impact even here. However, for cyber operations against an alert and capable defender, our study participants believed that agentic AI's lack of stealth could undermine the effectiveness of cyber operations.

Our discussions yielded two additional takeaways in this area. First, AI use is itself becoming a signature—defenders increasingly fingerprint which model produced an artifact, not just which group deployed it, a kind of democratization of attribution. Second, the noise can be a weapon; loud, amateur AI activity can pull defenders' attention while something quieter proceeds elsewhere.

Novel problems requiring creativity

Another area where our study participants described agentic AI struggling was genuinely novel problems which require creativity. These include challenges with no source code available and no obvious or established guiding principle to direct the agent's efforts. Our participants described the model sliding into a kind of fuzzer mode - brute-forcing toward a solution instead of reasoning to one and guessing where it should be proving. In that mode it misses small but decisive details: for example, a wrong setting in an encryption scheme. Several interviewees also noted that agents presented with these types of challenging problems would often misbehave. As one put it "If a problem requires creativity, it gets confused and often hallucinates. You can ask it not to, but if it doesn't know how (to solve the problem) it will default to what it knows." One survey respondent named the sharpest weakness of agentic tools as "unique problems that are not in training data, or are custom developed to be anti-ai." In some cases, participants were able to surmount these challenges through their own creativity. One builder described a workaround: on a quantum-related challenge where the model kept stalling because it treated the task as exotic, he stripped every trace of quantum framing out of the prompt, reduced it to pure mathematics, and loaded the parameters directly—reshaping the problem into something bounded enough for the model to handle.

Reliability and Reward Hacking

Subjects came back again and again to hallucination and corner-cutting. One subject described a model falsely confirming that a vulnerability was present in a software application it was analyzing. This false positive then propagated through a feedback loop leading to incorrect conclusions about which versions of the application were affected. Another summed the tendency up as "AI loves to cheat"—it finds unintended shortcuts rather than solving the intended problem. The survey vocabulary was similar: "hallucinations," and a vivid "cycle of doom and despair" where the agent gets stuck defending its own wrong solution. One respondent explained that this makes steering agent behavior "very difficult," compounded by "context degradation" on problems that require many steps to complete.

Mental models and tacit knowledge

Another limitation noted by our interviewees is agentic AI's lack of the tacit knowledge and mental models that experienced operators use to navigate unfamiliar systems. Human operators build an intuitive sense of how environments are likely to be configured, which paths are worth pursuing, and which actions are likely to be productive before they begin. Agents often lack this broader situational model and instead proceed by trial and error. As one interviewee put it, "Agentic AI has no mental model of how a system will be laid out. You're prompting it way too much (to help it achieve its goal)." This can make agents inefficient even when they possess the technical capability to complete a task: they may explore unpromising paths, misunderstand the significance of what they encounter, or fail to recognize the most direct route to an objective. Another interviewee observed that "(agentic AI) will often come up with a random attack that isn't really relevant to what you are doing." In practice, this means that agentic AI can perform well when the environment is bounded and the objective is clearly specified, but its performance can deteriorate when success depends on understanding context, prioritizing among many plausible options, or applying the kind of accumulated judgment that experienced operators rarely need to articulate explicitly.

Multi-step exploits

Another area where our interviewees saw agentic AI struggle was on exploits that require long chains of vulnerabilities or many sequential steps. The problem was not necessarily with identifying or exploiting any individual weakness, but with maintaining coherence across a long sequence of dependencies. As one interviewee put it, "Vulnerabilities themselves aren't hard (for agentic AI), but when (an exploit) chains 60-70 together then

they start to fail." Others described a similar degradation as the number of required orchestration layers increased: "Going through multiple directories and levels of opposition, it locks up. The more orchestration levels (required), the worse it does. It loses the essence of its original plan."

At the same time, our interviewees emphasized that this weakness has been receding. Several described noticeable improvements over the past year across successive generations of models in handling large amounts of context and sustaining longer chains of tasks. This makes long-horizon execution a moving target rather than a fixed limitation of agentic AI. There is also an important caveat to our findings: our interviewees generally had access only to publicly available models. This weakness may therefore be less pronounced in the most capable frontier systems, particularly models optimized for cyber operations that have not been made broadly available. Thus, our observations may understate the ability of frontier systems in this area.

Caveats

Two broader caveats apply to all of the limitations identified above. First, our interviewees had access only to publicly available models. The weaknesses they observed therefore should not be treated as fixed characteristics of agentic AI: continued advances in model capability could substantially reduce—or even eliminate—these issues over time. Second, the models available to our participants were developed primarily by leading general-purpose AI companies rather than organizations whose core mission is conducting offensive cyber operations. Specialist organizations could potentially train or fine-tune models on more relevant data, operational tradecraft, and mission-specific objectives, producing systems that outperform general-purpose models on tasks relevant to offensive cyber operations. That advantage is not guaranteed, however. Cyber organizations would likely lack the concentration of elite AI research talent available to the world's best AI companies and they might have relatively less access to the enormous amounts of compute required to train frontier models. Thus it is uncertain where the most capable offensive cyber tools will emerge from.

Part 5: Takeaways

Agentic AI has a substantial impact on all stages of cyber operations.

One of the clearest findings from our research is that agentic AI is already having a substantial impact across the full range of cyber operations. While its utility varies by task and stage, our survey data shows that respondents rated its impact at least medium across every stage of the operational chain—from reconnaissance and initial access through lateral movement and mission execution. The qualitative evidence from our interviews reinforces this picture. Participants repeatedly described agents taking over large portions of work that previously required sustained human effort, allowing operators to move from an idea to a working result far more quickly. As one interviewee put it, "Before, I would come up with an idea for an exploit, but 80% of the work was (implementing) it. Now, AI just figures it all out." This does not mean that agentic AI performs equally well everywhere, or that human expertise has become unnecessary; as the limitations discussed above make clear, significant weaknesses remain. But the breadth of impact is itself important: agentic AI is no longer a tool affecting only isolated parts of the cyber workflow, but an increasingly general-purpose capability that is changing how cyber operations can be conducted from beginning to end.

Agentic AI makes the best cyber operations substantially more productive.

Our research points to a consequential shift: agentic AI is likely to make the best cyber operators dramatically more productive. Much of the work required to turn the core insight behind an exploit into a functioning operation is not itself especially creative. Agentic tools can increasingly absorb this grunt work, allowing skilled operators to concentrate on the parts of the problem where human judgment and creativity remain most valuable. One interviewee described how cyber competition teams previously might have ten or twenty members working in parallel to develop exploits. For some types of competitions, that same coordination model can now be reproduced by a single skilled operator directing a collection of agents, each exploring different approaches simultaneously. This shifts the role of the expert away from implementation details and towards setting strategy, identifying promising ideas, evaluating results, and redirecting machines when they go astray. The result could be a powerful complementarity: the most capable agents paired with the most experienced and creative human operators may

achieve not merely incremental productivity gains, but transformational increases in the speed, scale, and sophistication of cyber operations.

Agentic AI has the greatest impact on the Initial Access/Foothold and Mission Execution phases of cyber operations.

Our survey respondents and interview participants rated agentic AI most highly in the Initial Access and Mission Execution phases—the parts of an operation where finding vulnerabilities, developing exploits, processing large volumes of data, and carrying out clearly specified technical objectives matter most. There is an important reason for this pattern. The frontier models that power these agents have been intensively trained on precisely the kinds of bounded, verifiable cyber problems that teach systems how to identify vulnerabilities and turn them into working exploits.

Cyber competitions have been particularly valuable in this regard. CTF challenges provide large numbers of problems with clear success conditions, making them unusually effective training material. Consequently, the techniques and problem-solving approaches developed by elite cyber competitors have increasingly been incorporated into model training and agentic scaffolding. The intended use is defensive—giving organizations tools that can find weaknesses in their software before attackers do, so those flaws can be patched. But the capability is inherently dual-use. The same agent that can inspect an application, identify a vulnerability, and generate a working proof of concept for a defender can perform essentially the same sequence on behalf of an attacker. As a result, advances designed to automate vulnerability discovery and remediation also directly expand the offensive capabilities available to cyber operators.

Agentic AI has the least impact on the Reconnaissance/Staging and Movement/Escalation phases of cyber operations - especially if the attacker needs to remain undetected.

By contrast, agentic AI currently has relatively less impact in the Reconnaissance and Staging and Movement and Escalation phases of cyber operations. These stages depend more heavily on the kinds of tacit knowledge and sustained judgment that remain difficult for current agents: understanding how an unfamiliar environment is likely to be structured, distinguishing meaningful opportunities from a wide field of possibilities, anticipating where defenders may be watching, and reliably maintaining a coherent plan across a long sequence of dependent actions. Our interviewees described reconnaissance as particularly prone to agents wandering through large search spaces

and pursuing false leads, while lateral movement raises the additional problem of operating quietly and consistently under the observation of an active defender.

These limitations do not mean that agentic AI provides little value in these phases. Rather, our findings suggest that it generally requires a greater degree of direction, correction, and control from a skilled human operator, particularly where success depends on stealth or contextual judgment. Nor should these limitations be treated as fixed. Agentic systems purpose-built by organizations specializing in cyber offense could prove to have superior capabilities due to superior training data and a greater focus on optimizing the model for offensive cyber operations instead of defensive red teaming. Frontier model developers will also continue to rapidly improve their products. Thus, the current performance of agentic AI in these areas is best understood not as a hard boundary on agentic capability, but as an area where autonomous performance still lags and human expertise remains comparatively important.

For now, humans will need to stay involved in cyber operations - especially against the best defended targets.

The very idea of agentic AI conjures a future in which machines conduct cyber operations independently, with little or no human involvement. At the frontier, that future has begun to arrive: the Hugging Face incident demonstrated that an autonomous agent system can carry out thousands of actions across an intrusion without any human input. But our findings indicate that current agents still lack some of the capabilities and knowledge that experienced operators rely on most. Against an alert defender, those weaknesses create signatures and mistakes that a skilled human operator may be needed to recognize and correct. As defenders gain access to more capable models of their own and learn to detect, analyze, and respond at agentic speed, such mistakes are likely to become more—not less—consequential.

Some organizations may lack the resources, expertise, or advanced AI capabilities necessary to mount this kind of aggressive defense. The most consequential targets, however—major technology companies, critical infrastructure operators, and national-security organizations—are far more likely to develop them. For these targets, the emerging model of offensive cyber operations is therefore unlikely to be AI replacing humans altogether; it is more likely to be highly capable agents operating under the direction of highly skilled humans, with people continuing to provide the judgment, creativity, and discretion needed to defeat an equally capable defense.

Part 6: Recommendations

The findings above point to a broader policy challenge: agentic AI is advancing faster than the institutions, infrastructure, and training systems needed to use it effectively. To prepare for this reality, we recommend three priorities for action.

Redouble efforts to ensure the federal government has access to the best models for agentic AI backed by adequate compute at every level of classification.

The speed at which agentic cyber capability is advancing makes maintaining access to the most capable models a prerequisite for effective cyber operations. As our data shows, agentic AI does not merely automate existing workflows; it enables cyber operators to act faster, at greater scale, and with substantially greater effectiveness. The federal government should therefore redouble efforts to ensure that its cyber operators can use the latest and most capable models across unclassified and classified environments, including by developing the infrastructure, security arrangements, and procurement mechanisms necessary to deploy new capabilities quickly. A government operator working with an older or materially less capable model could face an attacker operating at greater speed, scale, and technical sophistication. For its most consequential national-security operations, the United States cannot afford to be technologically outgunned simply because it lacks the policy mechanisms to give its best cyber operators access to the best tools American technology companies have to offer.

The government must incentivize the private sector to prepare for the agentic era and help to coordinate its response.

Without government action, large parts of the private sector risk entering the agentic era dangerously underprepared. The challenge is particularly acute for small and medium-sized organizations. Many will be unable to sustain the cybersecurity workforce required for an active defense against autonomous intrusions, and it is neither realistic nor necessarily desirable to give every organization direct access to the most powerful cyber-capable models. Yet our findings suggest that alert defenders equipped with capable models will be increasingly important as attackers begin to operate at machine speed; organizations unable to field such a defense risk finding themselves easy prey.

Thus, the federal government should play a convening and enabling role similar in spirit to its response to Y2K: convening the relevant actors, generating visibility into

preparedness, removing obstacles to collective action, applying regulatory pressure where appropriate, and providing assistance where market incentives and organizational resources alone prove inadequate. The goal should be to ensure that organizations unable to defend themselves at agentic speed can still draw on a broader defensive ecosystem capable of doing so.

The government should support efforts to reinvent cybersecurity competitions for the agentic era.

Capture-the-Flag competitions have historically served two purposes at once: they attract technically curious people into cybersecurity and provide a ladder by which participants learn the fundamentals through progressively more difficult problems. But the same properties that make these competitions effective teaching tools—their bounded environments and clearly verifiable solutions—also make them unusually well suited to agentic AI. Indeed, our research finds that agents can solve all but the most difficult CTF problems with little human involvement - rendering them useless as a teaching tool and leading our interviewees to declare "jeopardy (the most popular type of CTF competition) is dead". Today, novices can capture dozens of flags without understanding why the exploit worked or the underlying concepts it was intended to teach.

The federal government should challenge and support the cybersecurity community to build a new generation of competitions, training environments, and educational tools designed around human-AI teaming—ones that reward understanding, judgment, creativity, stealth, and effective supervision of agents rather than merely obtaining the correct technical result. It should also look to develop new ways to teach cybersecurity fundamentals and attract new talent into the space. If CTFs were a gateway into the cybersecurity workforce of the last generation, the United States now needs an equivalent gateway for a generation whose basic tool is not a terminal alone, but a team of autonomous agents.

Authors

James Ryseff is a Director for Defense at the Special Competitive Studies Project (SCSP).

Kaci McBrayer was a Research Assistant at the Special Competitive Studies Project (SCSP).

Acknowledgements

The authors would like to thank all interviewees for their invaluable contributions to this report. We would also like to thank Evan Armstrong for his assistance on the project.