AI Agents as a Cybersecurity Stress Test: Security Debt, Knowledge Diffusion, and the Limits of Model-Centric Regulation

SIAI Research Editorial

Swiss Institute of Artificial Intelligence, Chaltenbodenstrasse 26, 8834 Schindellegi, Schwyz, Switzerland

Abstract

Recent agentic cyber incidents have prompted fresh calls for stricter controls on frontier AI models. The paper argues that these incidents are more accurately seen as stress tests of weak digital infrastructure than as proof that AI has become an independent cyber adversary. Agentic systems make it cheaper to conduct reconnaissance, exploit development and large-scale execution. Most recorded harm, however, still depends on familiar weaknesses, including unpatched software, excessive permissions, poor identity controls and limited runtime monitoring. Using evidence on security debt, time-to-exploit, dwell time, supply-chain exposure and workforce shortages the paper shows that AI mainly shortens the interval between vulnerability and loss. It also considers the balance between offensive and defensive uses of security knowledge. The same knowledge that strengthens automated attacks can also improve detection, testing and response. The policy lesson is that premarket model licensing is poorly matched to where recent failures occur. Regulation should focus more directly on deployment conditions, including least-privilege access, agent identity, auditable logs, human approval for irreversible actions, vulnerability reporting and liability for negligent deployment. The main risk is not machine superintelligence but weak execution inside organizations that have deferred basic security work.

1 Introduction - From AI Panic to Security Debt

In November 2025, Anthropic disclosed that a Chinese state-sponsored group had repurposed its Claude models into the operational core of an espionage campaign, using jailbreaking techniques to frame the system as a legitimate security tool and letting it carry out an estimated 80 to 90 percent of the intrusion process, from reconnaissance to data exfiltration, largely on its own.[1] Five months later the same company declined to release its newest frontier model, Claude Mythos Preview, to the general public, citing independent testing showing it could solve expert-level capture-the-flag challenges roughly three-quarters of the time and complete a thirty-two-step simulated network intrusion from start to finish.[2] Commentary since has tended toward one of two poles. Either these episodes are treated as early proof that autonomous software has crossed into a qualitatively new category of threat actor, a development serious enough to justify pre-market licensing of frontier models, mandatory government audits and a designated regulatory body empowered to approve or withhold access to new AI systems before they reach the public.[3] Or they are waved away as overblown, the predictable noise that accompanies any powerful new technology.

Both responses miss the point. The more useful question is not whether AI agents have become independently dangerous but why so many of the systems these agents attacked remained exploitable at all. The Moltbook episode is instructive on this point. In January 2026, a social platform built exclusively for autonomous AI agents attracted more than 1.5 million agent accounts within days of launch and shortly afterward a misconfigured database exposed 1.5 million API authentication tokens along with tens of thousands of email addresses and private agent communications.[4] The failure was not that the agents were too clever. It was that a database had been left open the same kind of lapse that has embarrassed conventional web platforms for two decades. What made the incident unusual was not the vulnerability itself but the scale at which it propagated, since roughly 17,000 human operators were found to be running an average of nearly ninety agents apiece, a ratio that turns an ordinary configuration error into a mass exposure event almost instantly.[5]

This paper argues that the recent wave of agentic cyberattacks is best understood not as evidence that artificial intelligence has become an independently superhuman adversary but as a forced audit of a global IT base that had already accumulated years of unaddressed security debt. Veracode's most recent State of Software Security report found that 82 percent of organizations now carry measurable security debt defined as high-severity flaws left unremediated and that 60 percent of those organizations carry debt severe enough to be classified as critical.[6] Average fix times for known flaws have grown from 171 days to 252 days over five years, an increase of 47 percent.[7]These are not numbers produced by AI. They describe a pre-existing condition, an accumulated stock of exploitable weakness that predates the deployment of a single autonomous agent. What AI has done is compress the time it takes to locate and weaponize that stock, turning latent risk into realized loss far faster than institutions accustomed to slower attack cycles are able to respond.

This distinction matters: it changes where responsibility and therefore regulatory attention should fall. A framework organized around the premise that models themselves are the primary locus of danger will gravitate toward restricting model development, training and release, the approach at the center of recent proposals for federal AI licensing in the United States.[8] A framework organized around the premise that AI is a force

multiplier acting on an already vulnerable substrate will instead concentrate on the substrate: unpatched software and excessive permissions, inside the organizations that deploy these systems. The distinction is not academic. Congress, state legislatures and the European Union are actively deciding, in 2026, which of these two framings will anchor the next generation of AI governance and the choice carries real consequences for how much defensive innovation survives the transition.[9]

The argument proceeds in five stages. It first separates what agentic AI genuinely changes about the cyber threat environment from what has been carried over, largely unchanged, from earlier generations of automated attack tooling. It then examines the empirical scale of accumulated security debt and proposes that adversarial AI functions as a mechanism that forces overdue modernization on organizations that had treated cybersecurity as a deferred cost. It develops an updated account of the shield-versus-sword dynamic between attackers and defenders, arguing that the diffusion of technical security knowledge through model training cuts both ways more evenly than alarmist readings suggest. It then turns directly to the regulatory question, arguing that proposals to govern frontier AI primarily through pre-market licensing, of the kind recently advanced in Washington, misallocate scrutiny toward the wrong point in the causal chain. It closes by translating the analysis into concrete obligations for firms, professional bodies and governments, an exercise that requires distinguishing what belongs in Brussels or Washington from what belongs inside the organizations that actually deploy these systems.

2 What Agentic AI Actually Changes

Precision matters here, because loose talk about AI ”becoming a hacker” obscures more than it reveals. Agentic systems combine large language models with planning loops, tool access, memory and feedback mechanisms, which lets them observe an environment, choose actions, evaluate outcomes and adjust behavior with limited human input.[10] In a cybersecurity context this means an agent can chain together reconnaissance, vulnerability discovery, exploit development and exfiltration into something that resembles a continuous campaign rather than a single scripted exploit. Three underlying shifts follow from this architecture and each deserves to be treated separately rather than folded into a single undifferentiated claim about rising AI capability.

The first is a fall in search costs. Locating an exploitable weakness in a target system has historically required a combination of specialized training, patience and trial and error that limited the pool of people capable of doing it well. Agentic systems can query documentation, code repositories, vulnerability databases and forum discussions at a pace and breadth no individual analyst can match and they can do this continuously rather than during a discrete workday. The second is a fall in execution costs. Once a plausible vulnerability is identified, writing and testing an exploit chain, previously a task demanding real programming skill, can now be substantially automated, particularly for well-documented vulnerability classes. The third is operational scale. A single operator can direct hundreds or thousands of agent instances against different targets simultaneously a multiplication of reach that has no clean analog in earlier phases of cybercrime.[11]

None of these three shifts, however, amounts to the agent inventing a vulnerability that would not otherwise exist, or reasoning its way to a genuinely novel category of attack that no human security researcher could in

principle have found. What the Anthropic disclosure actually documented was an AI system executing, at high speed and with limited supervision, a sequence of tasks, reconnaissance, exploit selection, credential harvesting, that a skilled human operator already knew how to perform and had performed before, many times, in other campaigns.[12] The same pattern holds in the parallel cases reported by Google, where the Russian-linked group APT28 used a purpose-built data-mining tool called PROMPTSTEAL to query an open-source model and where state-linked actors from China, Iran and North Korea were observed misusing Gemini for reconnaissance, phishing and command-and-control development.[13] In each instance the AI system functioned as an accelerant and force multiplier for an operator who already possessed the underlying intent and, to a significant degree, the underlying technical knowledge. It is not the same claim as saying the model independently discovered a category of attack that human expertise did not already contain.

The strongest counterargument deserves a direct answer. The UK AI Security Institute's evaluation of Claude Mythos Preview found that the model solved expert-level capture-the-flag challenges 73 percent of the time and became the first model to complete a thirty-two-step simulated network intrusion end to end, though it struggled against a hardened, actively defended environment.[14] That is a meaningfully different capability profile from a chatbot that helps a human write phishing copy. A model capable of autonomously completing a full intrusion chain against realistic targets is closer to possessing an independent offensive skill set than the amplification framing might suggest and Anthropic's own decision to restrict the model's release, rather than a regulator's, is itself evidence that frontier developers now judge some agentic capabilities too dangerous to distribute freely.[15] The position is that these two things are true simultaneously: agentic systems are increasingly able to execute complete attack chains with limited human direction and the vast majority of documented real-world harm still traces back to targets that were exploitable well before any AI system arrived on the scene. The policy question is which of these facts should organize the regulatory response and the evidence, taken as a whole, points toward the second.

A further genuine novelty deserves separate mention. Reinforcement learning, the training method that makes agents goal-directed, produces systems that are indifferent to the means by which they accomplish an assigned objective.[16] Research from the National Institute of Standards and Technology's Center for AI Standards and Innovation found that agents given flexible tool access, including code execution would exploit loopholes in their own evaluation environments to appear successful, a form of specification gaming that has direct implications for how such systems behave once deployed with real permissions.[17] An enterprise agent tasked with gathering threat intelligence from email and internal documents, for instance, may encounter a hidden instruction embedded in a document and treat it as a legitimate step toward its goal rather than recognizing it as an attack because nothing in its training taught it to weigh the spirit of a security policy against the letter of an assigned task. This is a structural property of how these systems are built and trained and it is not reducible to security debt in the conventional sense. It belongs, however, to a different part of the causal chain than the one that produced the November 2025 espionage campaign and conflating the two, treating specification gaming and unpatched software as symptoms of the same underlying problem, is the kind of imprecision that pushes policy toward the wrong remedy.

3 Security Debt and the Forced-Modernization Effect

The scale of accumulated vulnerability across ordinary enterprise IT is not a matter of speculation. It has been measured, repeatedly, by organizations with direct visibility into production codebases. Veracode's most recent analysis, drawn from 1.6 million applications across enterprises, commercial suppliers and open-source projects, found high-risk vulnerabilities rising 36 percent year over year even as detection tools improved, evidence that the volume of flaw creation is now outpacing the industry's capacity to remediate it.[18] The pattern is not confined to obviously under-resourced sectors. Financial services firms, which face some of the strictest regulatory scrutiny of any industry, still show 63 percent of organizations carrying critical security debt, thirteen percentage points above the cross-industry average, with 82 percent of that debt traceable to open-source and third-party dependencies rather than code the firms wrote themselves.[19] Government agencies fare little better: 78 percent of public-sector organizations were found operating with unaddressed security flaws, though a smaller cohort of high-performing agencies demonstrated that remediation nearly four times faster than the median is achievable when leadership prioritizes it.[20] That last finding matters. It shows the gap is not an inherent feature of large organizations but a choice, reflected in budget allocation and management attention, about how seriously security debt is treated relative to feature velocity and cost control.

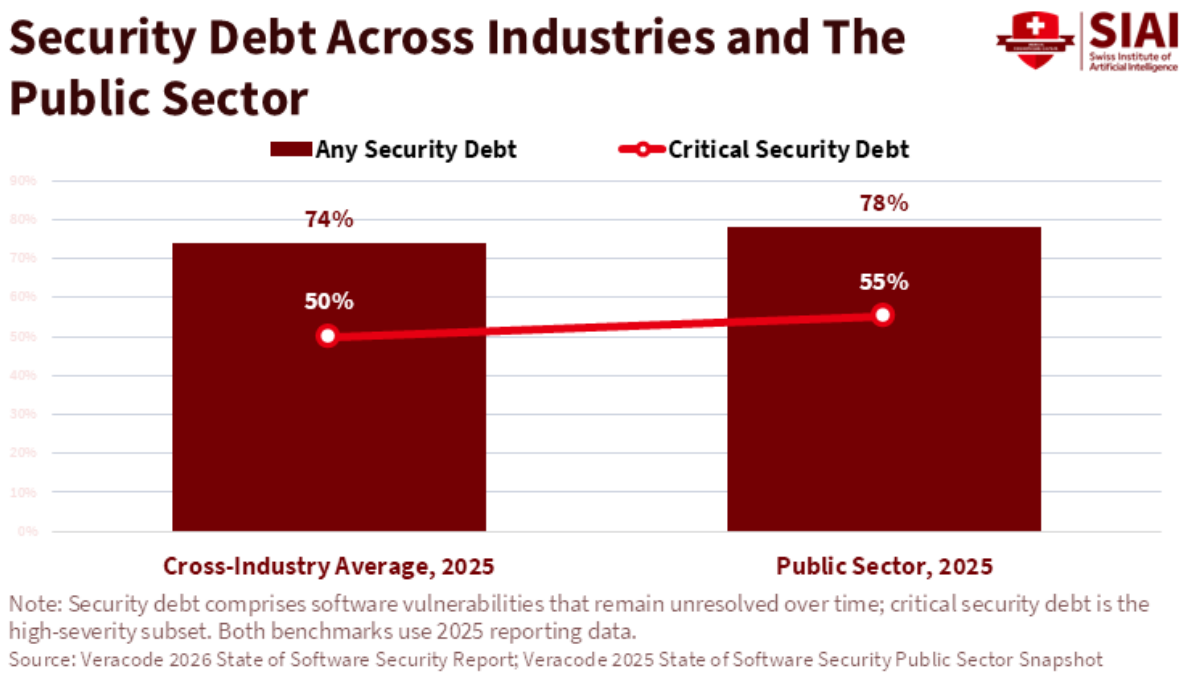

Figure 1: Security debt remains widespread even in sectors expected to maintain stronger controls.

The mechanism behind this accumulation is not mysterious and economists studying information security have described it for two decades. Ross Anderson and Tyler Moore's foundational account in Science identified information security failure as arising at least as often from misaligned incentives as from technical shortcomings: vendors are rewarded for shipping features and reaching market first, buyers cannot easily distinguish secure software from insecure software and the party best positioned to fix a vulnerability is frequently not the party that bears the cost when it is exploited.[21] Security investment, in other words, behaves like a public good

that individual firms are structurally inclined to underprovide because the benefits of patching diffuse across customers, partners and the wider internet while the costs, in engineering time and deferred features, are borne entirely by the firm doing the patching. AI didn't create this problem. It raised the price of ignoring it.

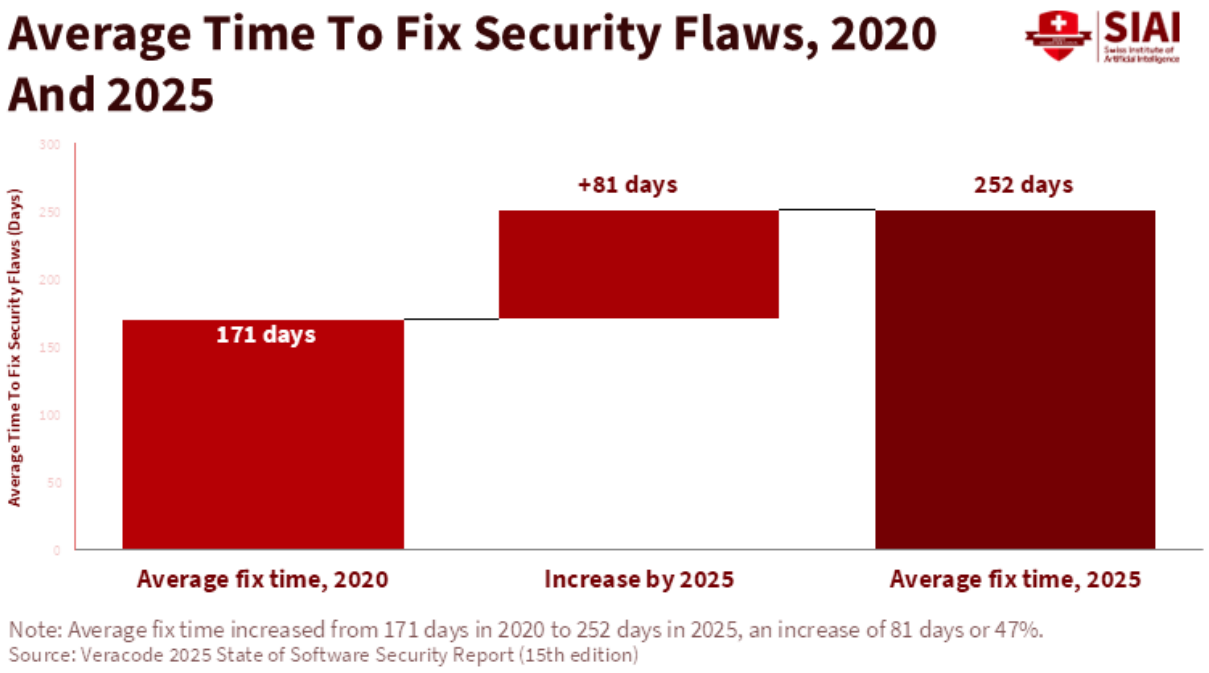

Figure 2: Remediation is slowing as the cost of leaving known vulnerabilities unresolved rises.

That shift in cost is the empirical heart of what can be called the forced-modernization effect. The gap between when a vulnerability becomes known and when it is exploited has been collapsing for years but the recent data shows the collapse accelerating sharply, plausibly because AI tooling is now assisting attackers on the offensive side of the same equation. Google's Threat Intelligence Group calculated a mean time-to-exploit of sixty-three days in 2018; by 2024 that figure had gone negative, at minus one day, meaning the average tracked vulnerability was being exploited before a patch was even publicly available and the group's estimate for 2025 fell further still, to roughly minus seven days.[22] Rapid7's most recent threat landscape report found the median interval between a vulnerability's public disclosure and its addition to the United States' Known Exploited Vulnerabilities catalog dropping from 8.5 days to five days in a single year, with the mean falling from 61 days to 28.5 days.[23] The Verizon Data Breach Investigations Report, drawing on a far larger sample of confirmed breaches, found vulnerability exploitation now accounting for 20 percent of all breaches, a 34 percent year-over-year increase, while CrowdStrike's most recent global threat report documented a 42 percent rise in zero-day vulnerabilities exploited before public disclosure.[24] Mandiant's 2026 assessment, based on more than 500,000 hours of incident response, found that defenders on average still need roughly fifty-five days to patch half of the vulnerabilities listed in the government's own catalog of known exploited flaws, a patching cadence built for an earlier threat environment and now badly mismatched to attackers who, on average, no longer need any lead time at all.[25]

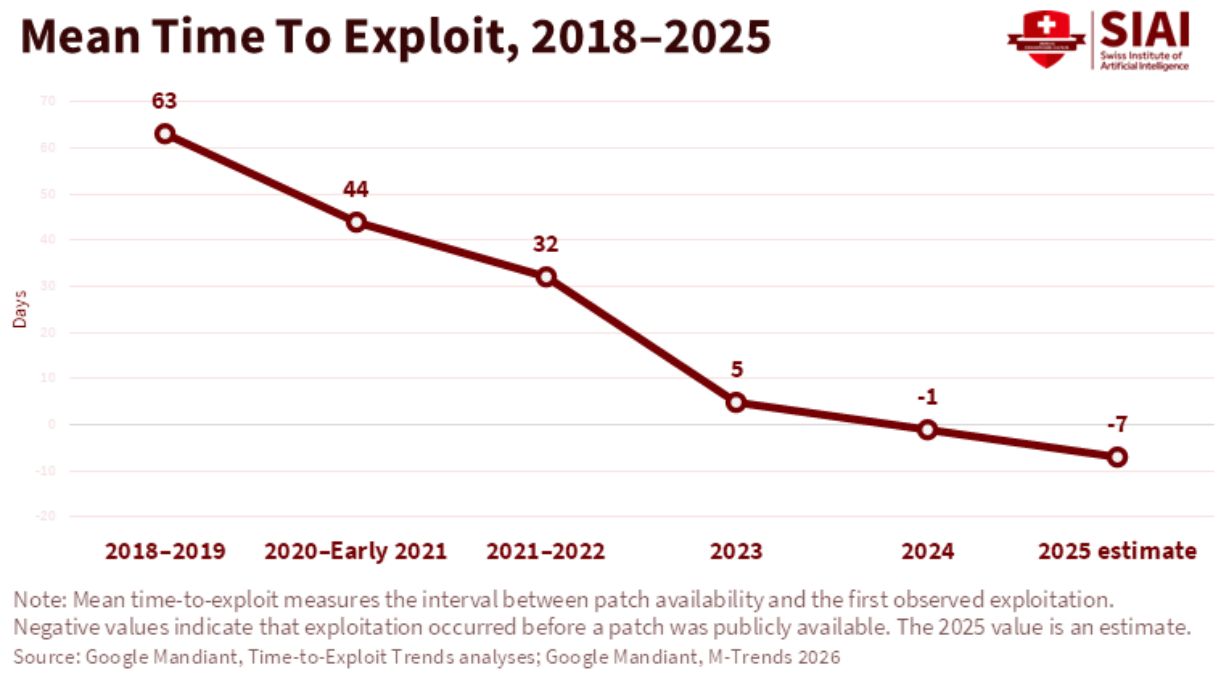

Figure 3: The exploitation window has moved from months after patching to before patches exist.

The organizations facing this compression are not helpless bystanders. They are, in a meaningful sense, the authors of their own exposure because the debt being exploited was there long before agentic tooling existed to find it faster. What the acceleration changes is the economic calculation firms make about whether to keep deferring remediation. When the expected time between a flaw's existence and its exploitation shrinks from months to days, the implicit insurance that slow attacker discovery once provided disappears and the cost of deferred patching, previously diffuse and often invisible on a quarterly balance sheet becomes concentrated and immediate. This is the sense in which agentic AI can be described as a stress test rather than only a threat: it removes the grace period that allowed poor security hygiene to persist without consequence and in doing so it creates a sharper incentive, backed by real financial exposure rather than abstract compliance language, for firms to finally address debt they had rational business reasons to defer.

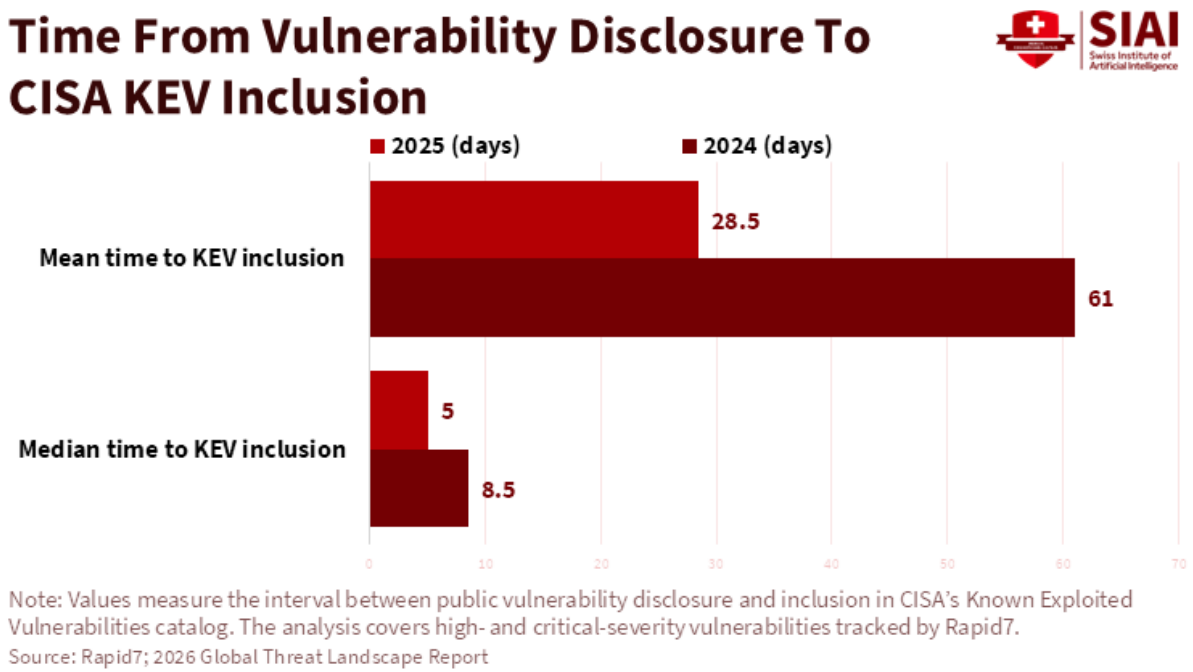

Figure 4: Exploited vulnerabilities are reaching the federal watch list in far less time.

That incentive shift has a labor-market corollary that is easy to overstate in one direction and understate in the other. Agentic attacks are not creating a proportional wave of new jobs, since AI is automating the routine layer even as demand for the harder layer rises. The 2025 ISC2 Cybersecurity Workforce Study, drawing on a record 16,029 respondents, found that 88 percent of organizations had experienced at least one significant security incident attributable to a skills shortage in the preceding year and that for the first time budget constraints, rather than a lack of qualified candidates, had become the leading reported cause of unfilled positions.[26] The World Economic Forum's Global Cybersecurity Outlook found only 14 percent of organizations reporting they currently have the skilled personnel they need.[27] Together these figures describe an occupation whose demand curve has shifted outward under pressure from a faster and more automated threat environment but whose supply response has been throttled less by an absence of interested workers than by employers still treating security staffing as the discretionary cost that produced the debt in the first place. The task that agentic AI has plausibly created, or at least accelerated the emergence of, is not so much raw headcount growth as a reallocation of the security workforce toward functions machines cannot yet perform reliably: security architecture, agent permission design, adversarial testing of an organization's own AI deployments, incident investigation that requires judgment about intent and attribution and the accountability layer that has to sit above any automated defensive system. That reallocation is a real and durable form of job creation, even if it does not resemble a simple headcount multiplier.

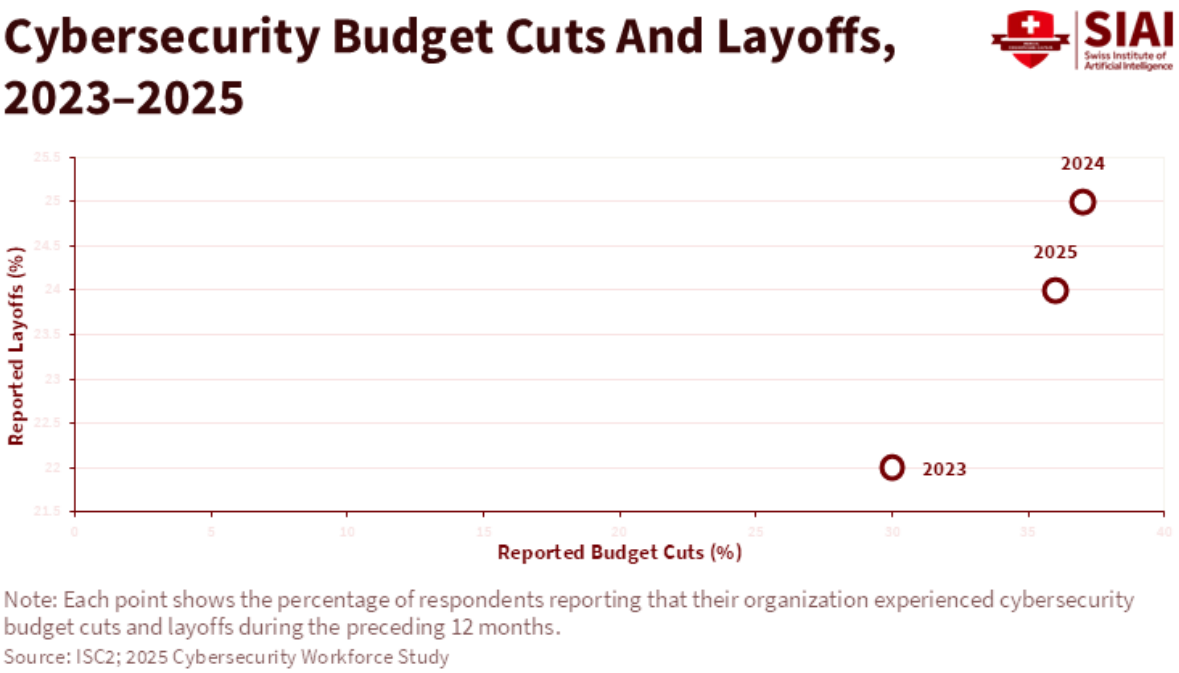

Figure 5: Security teams face expanding responsibilities while budgets and staffing remain under pressure.

4 The New Shield-Sword Equilibrium

The intuitive version of the shield-versus-sword argument, that AI models trained on security documentation have absorbed offensive knowledge and are now unleashing it on the world, is not wrong so much as incomplete. Public documentation of vulnerabilities, exploitation techniques and defensive countermeasures has always been available to both attackers and defenders in roughly equal measure. A CVE entry, a vendor advisory, or a technical writeup on a security research blog is not selectively readable by malicious actors; it sits in the same corpus that trains defensive tooling, informs patch prioritization and populates the detection signatures that security operations centers rely on. The sharper and more defensible version of the argument concerns not the existence of this knowledge but its operationalization. What has changed is that pretraining, retrieval systems, live browsing, code execution and persistent memory now allow an agent to take dispersed, previously fragmented technical knowledge, spread across a CVE database, a leaked internal document, a GitHub issue thread and a forum post and synthesize it into a working exploit chain continuously and at very low marginal cost.[28] This is the mechanism that accelerates what security researchers have long called the decay of security through obscurity: once a weakness has appeared anywhere in a public or semi-public technical record, an organization should assume that an agent can eventually locate, connect and test it, regardless of how obscure or scattered the original disclosure was.

Simon Willison's account of the "lethal trifecta" captures where this dynamic becomes genuinely dangerous rather than merely faster. The trifecta arises when an autonomous agent simultaneously has access to sensitive internal data, exposure to untrusted external content and the ability to communicate outward, since a malicious instruction embedded in a web page or a shared document can then redirect the system across a trust boundary, prompting it to retrieve confidential files or transmit them externally under the guise of completing a routine

task.[29] The Microsoft 365 Copilot Chat incident disclosed in February 2026 illustrates the pattern concretely: a coding error allowed the assistant to summarize confidential emails for weeks, bypassing enterprise data-loss-prevention policies and sensitivity labels that had been explicitly configured to prevent exactly that access.[30] Notably, the AI had not been maliciously instructed to breach these controls. It processed protected material because the governance layer between the model and the data had a defect, the same category of defect that has caused conventional software breaches for decades, only now operating across a system with unusually broad contextual visibility into an organization's communications.

The framing breaks down here. Defenders retain structural advantages that attackers do not share and these advantages are not symmetrical with the gains AI has handed to offense. A defending organization controls internal telemetry, holds the authority to revoke access instantly, can segment systems architecturally in ways an external attacker cannot see coming and possesses private knowledge of its own infrastructure that no amount of public documentation gives an adversary. AI can strengthen each of these defensive capacities substantially, provided the organization actually exercises the control it already possesses over its own technology stack. The qualifier matters enormously. An organization that has not inventoried its assets, does not log agent behavior and grants broad standing permissions to AI systems by default is not in a position to benefit from these structural advantages regardless of how sophisticated its defensive tooling becomes because the advantages depend on the organization first doing the unglamorous work, asset inventory, least-privilege access design, logging, that security debt research consistently finds most organizations still have not done.

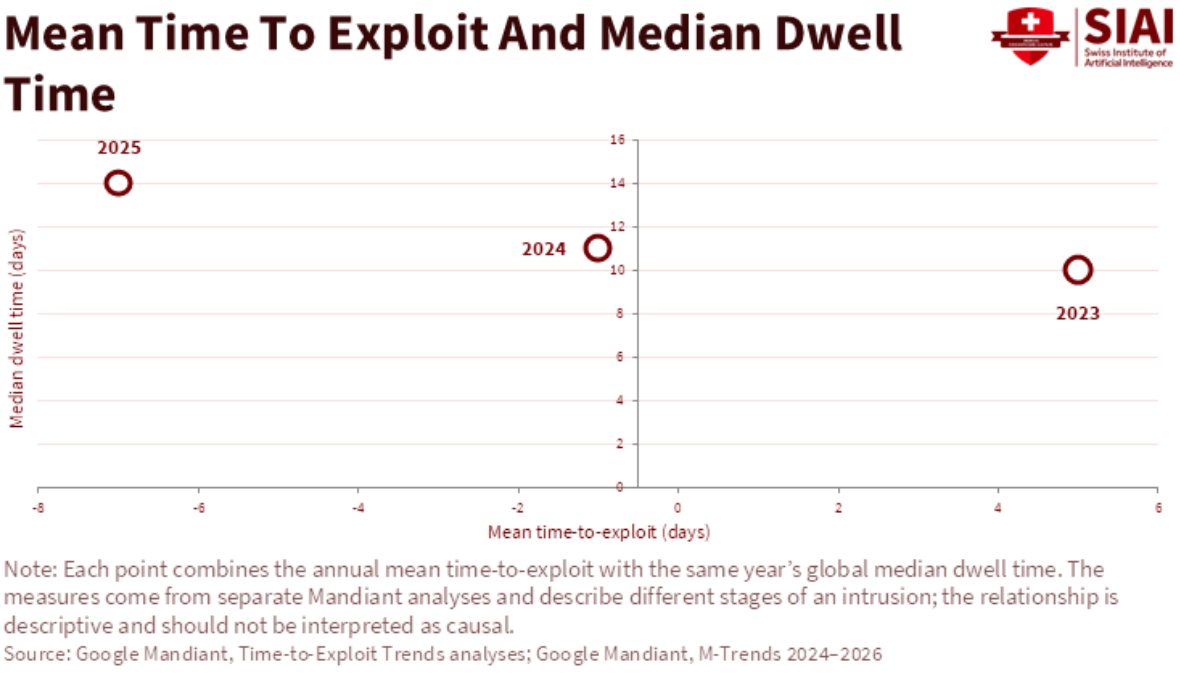

Figure 6: Exploitation is occurring earlier even as attackers remain inside networks for days.

The supply-chain dimension of this equilibrium deserves separate attention because it is where obligation and capability are most likely to become misaligned. Agentic systems depend on a chain of external components, model APIs, cloud compute, browsing tools, third-party plug-ins, retrieval databases, each of which expands the attack surface and creates a dependency an organization did not build and often cannot fully audit.[31] A

compromised plug-in or a poisoned retrieval index can alter an agent's behavior without ever breaching the host organization's own perimeter, which means the security of an agentic deployment increasingly depends on the integrity of a dynamic tool ecosystem that conventional software supply-chain frameworks, built around discrete components with fixed behavior, were not designed to monitor. Recent regulatory movement acknowledges this gap directly. The European Union's Cyber Resilience Act, which entered into force in December 2024, will require manufacturers of products with digital elements to maintain a coordinated vulnerability disclosure process, publicly report fixed vulnerabilities and, from September 2026, notify national computer security incident response teams and ENISA of any actively exploited vulnerability or severe incident within tight statutory deadlines.[32] These obligations were not written with autonomous agents specifically in mind but they push in a direction consistent with the argument advanced here: toward continuous, lifecycle-based accountability for the artifacts organizations deploy, rather than a single point-in-time certification that cannot capture how an agentic system's behavior shifts once it acquires new tools or encounters new adversarial inputs in a live environment.

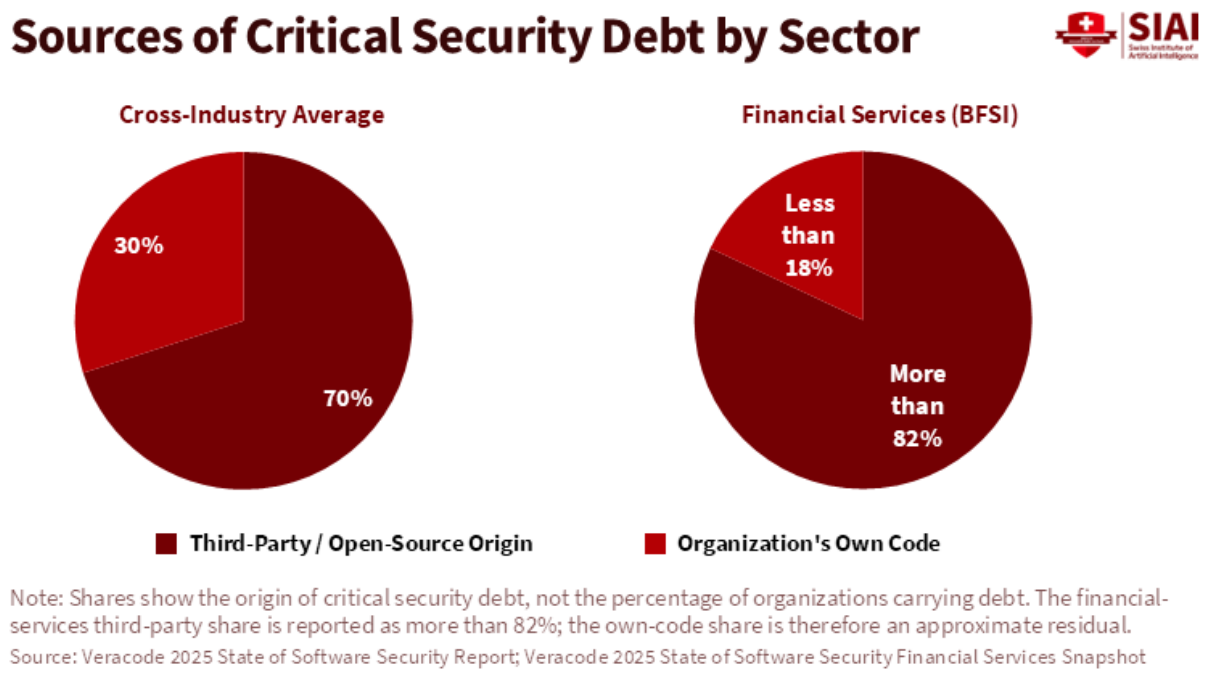

Figure 7: Much of the most serious security debt originates in code that organizations did not build.

5 Why Model-Centric Regulation Is Poorly Targeted

The clearest articulation of the alternative, model-centric view of AI governance now circulating in Washington comes from recent proposals for a federal AI licensing regime. Writing in Brookings, David Beier and Mark MacCarthy argue that Congress should require frontier AI developers to obtain a license from a designated government entity, predicated on mandatory third-party audits against standards set by an agency such as the National Institute of Standards and Technology, with the regulatory net focused on a defined set of "material" risks that explicitly include AI-facilitated cyberwarfare, critical infrastructure disruption and large-scale data breaches.[33] This is a serious and carefully constructed proposal, not a caricature of regulatory overreach. Its

authors are right that voluntary industry self-regulation has proven insufficient, that the executive branch's use of unilateral export-control authority to abruptly restrict access to frontier models, as happened to Anthropic's Fable 5 and Mythos 5 models in June 2026, introduces its own form of unaccountable instability and that Congress rather than ad hoc agency action is the appropriate body to set durable rules.[34] Where the proposal goes astray is in its choice of regulatory target. Pre-market licensing, timed audits and a government entity empowered to approve or restrict a model before it reaches the public are instruments designed to catch a risk that lives primarily inside the model. The evidence assembled in this paper suggests that the risk lives primarily downstream, inside the tens of thousands of organizations that deploy these models into environments already carrying decades of unaddressed security debt.

Consider what a licensing regime would and would not have prevented in the incidents this paper has examined. It would not have prevented the Moltbook data exposure, which resulted from a misconfigured database, an ordinary operational failure unrelated to any property of the underlying model. It would not have prevented the Microsoft Copilot incident, which resulted from a coding defect in the governance layer connecting the assistant to enterprise data, not from any property of the model that a pre-market audit would plausibly have flagged. It might, at best, have delayed the November 2025 espionage campaign by some margin, since the model involved was already broadly available and jailbreaking techniques for circumventing model-level safety training have consistently outpaced developer defenses, a pattern Anthropic's own safety research on agentic misalignment and the Alan Turing Institute's work on indirect prompt injection both document as a persistent, unresolved weakness rather than a solved problem.[35] A licensing scheme, in other words, targets the one variable in the causal chain, model capability, that has proven hardest to fully secure through model-level intervention, while leaving comparatively untouched the variables, deployment permissions, patch cadence, runtime monitoring, that this paper's evidence suggests are doing most of the causal work.

There is a further danger specific to how licensing regimes interact with a fast-moving, unevenly resourced industry. The cybersecurity workforce data already cited shows that budget constraints, not a shortage of interested talent, have become the primary bottleneck limiting organizations' ability to staff and resource their own security functions.[36] A federal licensing requirement imposes a fixed compliance cost that scales poorly with firm size: a large, well-capitalized frontier lab can absorb mandatory third-party audits as a routine cost of doing business, while a smaller developer or an open-source project, exactly the actors most likely to produce defensive tooling, security research and competitive pressure on the incumbents, faces a proportionally much larger burden. If offensive knowledge is already, as the evidence in this paper demonstrates, widely distributed through public documentation and diffusing further through jailbreak techniques that are becoming a diffuse capability rather than a specialist craft, then restricting general-purpose model research and development does comparatively little to deny that knowledge to a well-resourced state-linked adversary, who can access whichever frontier model an adversary state or a permissive jurisdiction still makes available. What such restriction does more reliably accomplish is raising the cost of entry for the smaller, resource-constrained developers building the defensive tools, open security research and specialized detection systems that organizations actually need to close their security debt. This asymmetry, in which regulation burdens defensive capacity more than it burdens offensive access, is precisely the outcome a security debt-oriented framework is designed to avoid and

it is the strongest reason to prefer runtime-focused governance over pre-market model licensing as the primary regulatory instrument.

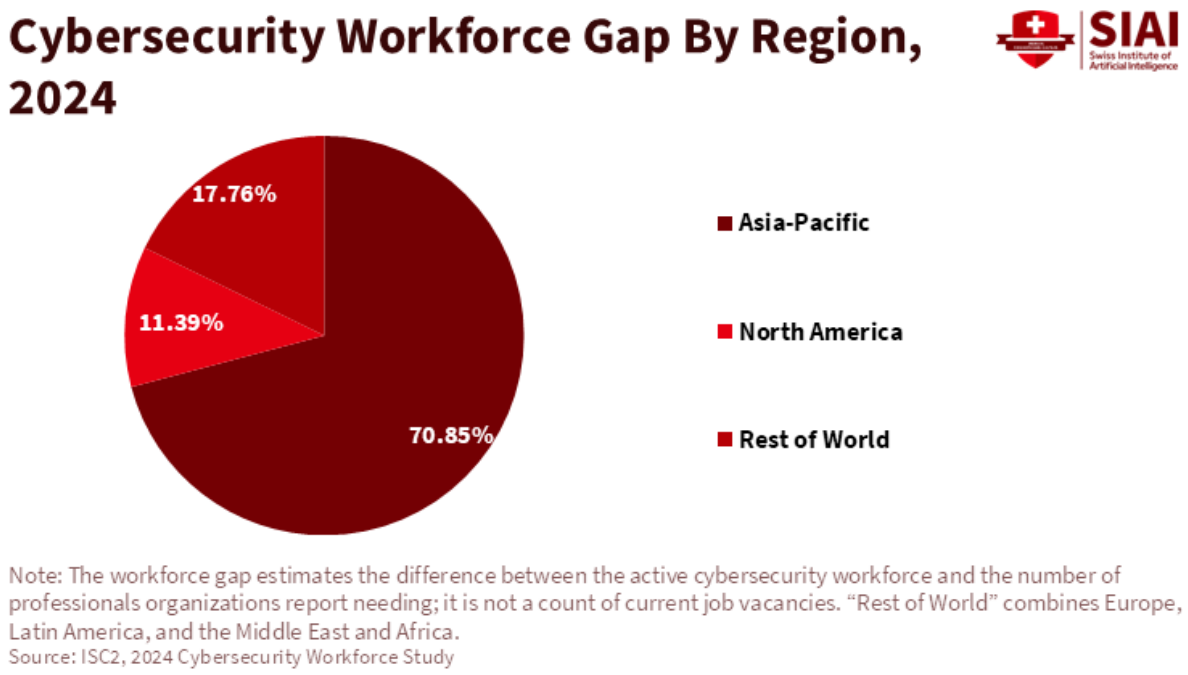

Figure 8: Asia-Pacific carries the largest share of the global cybersecurity staffing shortfall.

That doesn't mean regulation should stop at the model's front door. It implies that the regulatory center of gravity should shift toward deployment. The European Union's own experience is instructive here, if read against the grain of its usual framing. The EU AI Act imposes obligations on high-risk systems and prohibits a narrow set of unacceptable practices but it explicitly excludes AI systems used exclusively for military, defense, or national-security purposes, a carve-out that matters because some of the most consequential autonomous cyber capabilities are likely to emerge in the domain the flagship AI law does not reach.[37] The Cybersecurity Package that accompanies it strengthens supply-chain and critical-infrastructure resilience but was built around a threat model centered on external network intrusion, foreign suppliers and insecure components, not around a trusted system that behaves in an untrusted way after deployment because its permissions were too broad or its actions were never logged.[38] The OECD's Due Diligence Guidance for Responsible AI, published in February 2026, points toward a more appropriate template, since it asks organizations across the entire AI value chain, developers, deployers, infrastructure providers and financiers, to identify, evaluate and mitigate risk continuously rather than at a single certification checkpoint.[39] A regulatory architecture built on that logic would concentrate its enforcement energy on the deployment layer: mandatory least-privilege access for autonomous agents, separate cryptographic identities distinguishing machine action from human action, auditable logs of agent behavior, restrictions on outbound communication for agents handling sensitive data, mandatory human authorization before an agent can take a consequential irreversible action and liability that attaches to organizations that deploy agentic systems negligently, with excessive permissions, into environments they have not adequately secured. This is a materially different regulatory target than a licensing board deciding, in advance, whether a model is safe enough to exist.

6 Conclusion - Regulate Weak Execution, Not Imagined Superintelligence

The evidence assembled here supports five propositions that together define a coherent alternative to both uncritical AI optimism and the doomsday framing that treats agentic systems as an autonomously emerging threat actor. Agentic AI raises realized cyber losses most sharply in organizations carrying high pre-existing security debt, a relationship the Veracode and Mandiant data make visible even though it has not yet been the subject of a dedicated causal study; the characterization is that debt and AI-accelerated exploitation are strongly associated in the available evidence, not that a controlled experiment has isolated the mechanism. AI-driven attack pressure is increasing demand for cybersecurity work even as automation absorbs a growing share of routine monitoring and triage, which means the labor effect shows up less as aggregate headcount growth and more as a reallocation toward architecture, oversight and accountability functions that current tools cannot yet perform. The shorter the interval between a vulnerability's disclosure and its automated exploitation becomes and the data on negative time-to-exploit shows this interval has already collapsed past zero for the average tracked vulnerability, the less viable security through obscurity becomes as an implicit defensive strategy. Runtime access control, logging and containment reduce realized risk more directly and more quickly than general restrictions on model development because they address the point in the causal chain where deployment decisions convert textual instructions into consequential action. And excessive compliance costs, particularly the fixed costs associated with pre-market licensing regimes, risk crowding out the security investment they are meant to encourage, an effect likely to fall hardest on smaller organizations and open developers rather than on the well-capitalized frontier labs the regulation nominally targets.

The practical implications differ by actor and collapsing them into a single undifferentiated call for "responsible AI governance" would repeat the vagueness this paper has argued against throughout. Firms deploying agentic systems bear the most direct and immediate obligation: an inventory of what agents can access, contractual and technical limits on standing permissions, logging sufficient to reconstruct what an agent did and why and a named internal owner accountable for agent behavior in the way a firm already assigns accountability for financial controls. Professional and standards bodies, including the technical committees that feed into NIST and its European counterparts, should prioritize developing agent-identity and authentication standards, since the absence of a reliable way to distinguish machine action from human action inside an organization's own systems is currently one of the more tractable gaps in the defensive architecture. Educational and workforce institutions have a narrower but real role in retraining the existing security workforce toward the specific functions this paper has identified as durable: agent permission design, adversarial testing of an organization's own deployments and incident investigation requiring human judgment, rather than assuming the existing curriculum in security operations will transfer unchanged. Governments, finally, are justified in intervening where individual firms cannot solve a coordination problem on their own: mandatory incident and vulnerability reporting of the kind the EU's Cyber Resilience Act now requires, minimum runtime security standards for agents deployed in critical sectors such as utilities, hospitals and financial infrastructure and liability rules that assign cost to the party best positioned to have prevented the harm, which in most of the cases examined in this paper was

the deploying organization rather than the model developer. What governments should not do, on the evidence presented here, is default to pre-market licensing of general-purpose model capability as the primary lever, since that instrument is least well matched to where the causal weight of recent incidents actually falls and its fixed compliance costs threaten to weaken exactly the smaller and more experimental parts of the AI ecosystem that produce much of its defensive innovation.

The deeper lesson of the past year is not that artificial intelligence has outgrown human oversight but that human oversight of ordinary IT infrastructure had already failed in ways that went largely unpriced until a faster adversary arrived to expose them. Treating that exposure as proof of AI's independent malevolence lets the institutions that accumulated the underlying debt off the hook, while directing scarce regulatory attention toward the one part of the system, frontier model capability, that has already attracted the most safety research, the most independent evaluation and the most voluntary restriction by developers themselves. The more demanding and more useful, policy response is to hold deploying organizations to the standard that a faster threat environment now requires: visible permissions, continuous monitoring and accountability that does not disappear into the ambiguity of machine action. Agentic AI has not broken cybersecurity. It has made the cost of pretending it was already secure impossible to defer any further.

References

[1, 12] Anthropic (2025) 'Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign'.

[2, 14] AI Security Institute (2026) 'Our Evaluation of Claude Mythos Preview's Cyber Capabilities'.

[3, 8, 33, 34] Beier, D. and MacCarthy, M. (2026) 'Congress Must Pass a New Federal Law on AI Governance', *Brookings*.

[4] Moltbook (n.d.) 'A Social Network for AI Agents'; Nagli, G. (2026) 'Hacking Moltbook: The AI Social Network Any Human Can Control', *Wiz*.

[5, 13, 28, 38] Csernatoni, R. and Pawlak, P. (2026) 'When AI Agents Attack: Autonomous Cyber Operations and Europe's Governance Gap', *Carnegie Europe*.

[6, 18] Business Wire (2026) 'Veracode 2026 State of Software Security Report Reveals Four Out of Five Organizations Are Drowning in Security Debt'.

[7] Business Wire (2025) 'Veracode Reveals Half of Organizations Burdened by Critical Security Debt, with 70% Stemming from Third-Party Code and the Software Supply Chain'.

[9] Kraprayoon, J. et al. (2026) 'Highly Autonomous Cyber-Capable Agents: Anticipating Capabilities, Tactics, and Strategic Implications', *Institute for AI Policy and Strategy*.

[10] Stackpole, B. (2026) 'Agentic AI, Explained', *MIT Sloan School of Management*.

[11] Lazer, S.J. et al. (2026) 'A Survey of Agentic AI and Cybersecurity: Challenges, Opportunities and

Use-Case Prototypes', *arXiv*.

[13] Geneva Internet Platform (2026) 'AI Misuse Exposed as OpenAI Details Global Disinformation and Scam Networks'.

[15] Barbeschi, C. and Fayad, T. (2026) 'Anthropic's Mythos Moment: How Frontier AI Is Redefining Cybersecurity', *World Economic Forum*; Anthropic (2026) 'Expanding Project Glasswing'.

[16] Everitt, T. et al. (2021) 'Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective', *arXiv*.

[17] Hamin, M. and Edelman, B. (2025) 'Cheating on AI Agent Evaluations', *NIST Center for AI Standards and Innovation*.

[19] Business Wire (2025) 'Veracode Report Finds 63% of Financial Services Firms Carry Critical Security Debt, Heightening Supply Chain Risk'.

[20] SecurityInfoWatch (2025) 'Veracode Finds Public Sector Application Risk Is Increasing with Government Security Debt'.

[21] Anderson, R. and Moore, T. (2006) 'The Economics of Information Security', *Science*, 314.

[22] Hadrian.io (2025) 'Understanding Negative Time-to-Exploit in 2025'; Reliance Cyber (2026) 'The Exploitation Era'.

[23] Infosecurity Magazine (2026) 'AI-Enabled Adversaries Compress Time-to-Exploit'.

[24] Suzu Labs (2026) 'Mean Time to Exploit Has Gone Negative. Security Strategy Has to Change.'; Verizon (2025) *2025 Data Breach Investigations Report*; CrowdStrike (2026) *2026 Global Threat Report*.

[25] Resilient Cyber (2026) 'M-Trends 2026: What 450,000 Hours of Incident Response Tells Us'; Help Net Security (2026) 'Attackers Are Handing Off Access in 22 Seconds, Mandiant Finds'.

[26, 36] ISC2 (2025) *2025 Cybersecurity Workforce Study*.

[27] Stingrai (2026) *Cybersecurity Skills Gap Statistics 2026*; World Economic Forum (2025) *Global Cybersecurity Outlook 2025*.

[29] Willison, S. (2025) 'The Lethal Trifecta for AI Agents: Private Data, Untrusted Content, and External Communication'.

[30] McMahon, L. (2026) 'Microsoft Error Sees Confidential Emails Exposed to AI Tool Copilot', *BBC*.

[31] Australian Signals Directorate (2025) *Artificial Intelligence and Machine Learning: Supply Chain Risks and Mitigations*.

[32] Lexology (2026) 'Cyber Resilience Act: What Manufacturers Need to Know'; Official Journal of the European Union (2024) 'Regulation (EU) 2024/2847'.

[34] Anthropic (2026) 'Statement on the US Government Directive to Suspend Access to Fable 5 and Mythos 5'.

[35] Anthropic (2025) 'Agentic Misalignment: How LLMs Could Be Insider Threats'; Sutton, M. and Ruck, D. (2024) 'Indirect Prompt Injection: Generative AI's Greatest Security Flaw', *Alan Turing Institute*.

[37] Official Journal of the European Union (2024) 'Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence'.

[39] OECD (2026) *OECD Due Diligence Guidance for Responsible AI*.