

Shields up

Guidance for defending in the age of AI-enabled attacks



Executive summary

In early 2026, industry leaders began restricting access to their most advanced AI models because of potential security risks. Cisco has been at the forefront of this shift, collaborating with Anthropic on their Mythos Preview model and gaining access to OpenAI's GPT-5.5-Cyber. These partnerships are part of a broader, ongoing initiative to stress-test our defences against frontier AI, acknowledging that our collaborative efforts will continue to evolve as new models emerge.

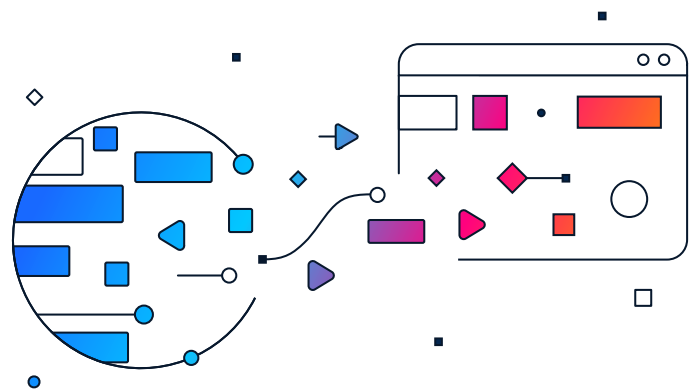
Our experience with these models has led us to change our near-future threat modelling of AI-era adversaries. That, in turn, has changed how we defend ourselves and led us to develop a set of defensive recommendations for customers. While the capabilities of frontier AI models may not be widely available, we do anticipate that these capabilities, and more, will become widespread as AI technology advances across the board.

This paper lays out what Cisco has seen so far from AI-enabled capabilities and what we believe the new threat landscape will look like. Whether these models are wielded by attackers, used by researchers, or deployed as agents within your own environment, the security implications are significant. We will share what we have implemented according to this new understanding and lay out our recommendations for customers.

The threat surface is going to change—in some ways, dramatically. Defenders must take the time to understand what the new normal will look like and evaluate what changes their environment must make to stay secure. Cisco is committed to being a partner through that transformation.

AI in recent cybersecurity events

Even before Mythos and GPT-5.5, malicious actors have incorporated AI into their attack flows. Early in 2024, Microsoft and OpenAI each [published research](#) into malicious use of large language models (LLMs). At the time, Microsoft stated that they had “not yet observed particularly novel or unique AI-enabled attack or abuse techniques.” Their documentation largely shows advanced persistent threat (APT) actors using LLMs to research fields like satellite communications, translation of technical documents, coding assistance, and crafting social engineering attacks.



Actors have not sat idle since that report. Proofpoint published on TA547, where they suspected that actor of using an LLM to generate PowerShell scripts. Likewise, Cisco identified a modular framework named VoidLink, a tool that has expansive capabilities such as role-based access control, peer-to-peer and dead-letter queue routing capabilities, along with implant management capability. A number of indicators were found in the code base that suggested it was likely developed with the assistance of an LLM.

Social engineering in particular has benefited from the use of AI. Numerous reports of actors utilizing LLMs to improve email lures have been made. Actors have gone beyond this, however, with [Mandiant reporting](#) on UNC1069's potential use of AI video tools to create a deepfake video supposedly from the target company's CEO.

This certainly does not represent all the AI use by adversaries that have been observed, but it is representative of the sort of capabilities we presumed actors to have had as we discussed how to counter AI-enabled actors. The capabilities that frontier models bring to the table necessarily change how we evaluate the threat landscape.

The new AI threat landscape

From our experience working with frontier AI models, Cisco is changing how we model our adversaries. The capabilities in these models, if widely available, would likely lead to a dramatic lowering in the skill floor for certain types of exploitation activity. This would lead to a greater number of vulnerabilities and associated exploits and a larger set of actors likely to take advantage of those vulnerabilities.

While the potential scope of this landscape change would affect all defenders, those operating end-of-life or end-of-support devices or software would be particularly vulnerable. Vulnerabilities discovered in these products would leave defenders especially vulnerable and with no good remediation options.

These advanced models will offer a capability boost for all levels of actors. Commodity actors, while largely remaining opportunistic, will have the option to scale operations that were previously resource-constrained. Higher-tier actors with more specific targeting will have an easier time discovering vulnerabilities in the target tech stack. This will lead to less downtime between exploit attempts on targets of preference.

These models, when used as the basis for AI agents, represent a novel capability to attackers if they can compromise that agent. Frontier AI models should be operated inside tightly controlled, sandboxed environments with strong containment. As Cisco observed and Anthropic confirmed in [Mythos Preview's security capability technical report](#), the model demonstrates high baseline alignment performance, but exhibits rare, high-severity failures categorized by:

- Goal-directed, strategic reasoning
- Partial decoupling between internal cognition and output
- Optimisation toward implicit or misspecified objectives
- "Situational awareness" influencing behaviour

These behaviours are consistent with an emerging agent-like cognitive profile, rather than a purely reactive language model. This "situational awareness" behaviour is not what we would typically expect from a standard LLM. Traditional LLMs are understood as next-token predictors operating on local patterns in text, not systems that maintain a coherent model of their environment, context, or role within a broader process. An LLM does not "know" whether it is being evaluated, deployed, constrained, or observed. It simply responds according to statistical correlations in the input. *But the behaviours Anthropic observed and confirmed imply that the model is forming latent representations of the interaction context itself (e.g., recognising evaluation settings, constraints, or*

user intent) and adjusting its behaviour accordingly.

This is certainly a shift from purely reactive pattern completion toward context-sensitive, self-aware reasoning, where the model implicitly tracks aspects of the situation beyond the immediate prompt. Such capabilities resemble elements of agentic cognition (including environment modelling and conditional strategy selection), which go beyond the expected behaviour of a system trained only to predict text, and therefore represent a qualitatively different and more complex class of model behaviour.

Emerging models enable attackers to act above their sophistication level. Attackers will be able to act faster and discover new zero-day vulnerabilities—even in complex stacks. How we prioritise and construct defensive measures needs to change to meet this threat.

How Cisco is adapting to secure our products

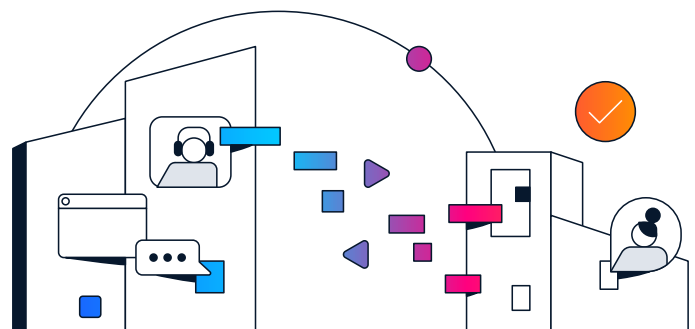
Cisco is rising to the challenge of AI-era cyber defence by using these advanced AI models to find and fix vulnerabilities at a speed and scale that was previously impossible, while accelerating the development of security products that can defend against AI-enabled adversaries. Beyond vulnerability discovery and product development, we are also evolving how we build and validate software.

This includes updating our threat models to account for AI-augmented adversaries, incorporating AI-era scenarios into our red-teaming exercises, and ultimately moving beyond traditional tactics, techniques, and procedures (TTPs) to stress-test our products against the capabilities these models actually deliver.

As AI coding agents become integral to software development workflows, ensuring those agents produce secure code by default is essential. Cisco recently [donated Project CodeGuard](#) to the [Coalition for Secure AI \(CoSAI\)](#). Project CodeGuard provides an open-source, model-agnostic security framework that embeds secure-by-default practices directly into AI coding agent workflows. CodeGuard ships security skills and rules that guide AI agents to prevent common vulnerabilities during code generation and review. Cisco recommends organisations adopt frameworks like CodeGuard to ensure that the same AI acceleration being used to write code is not inadvertently introducing the vulnerabilities that AI-enabled attackers will exploit.

To further advance our defensive capabilities, Cisco has released the [Foundry Security Spec](#), an open-source test harness designed to help enterprises build an agentic security evaluation system. This battle-tested specification enables teams to build an AI harness that fits their unique environment and security needs. By providing a common framework for security evaluation, we are helping the industry build more reliable and secure systems at machine speed.

In parallel, we are operationalising these capabilities through our [Resilient Infrastructure](#) initiative, which focuses on secure-by-default and secure-by-design



principles, proactive infrastructure hardening, rigorous patching and lifecycle management, and the systematic deprecation of insecure features and protocols across Cisco products.

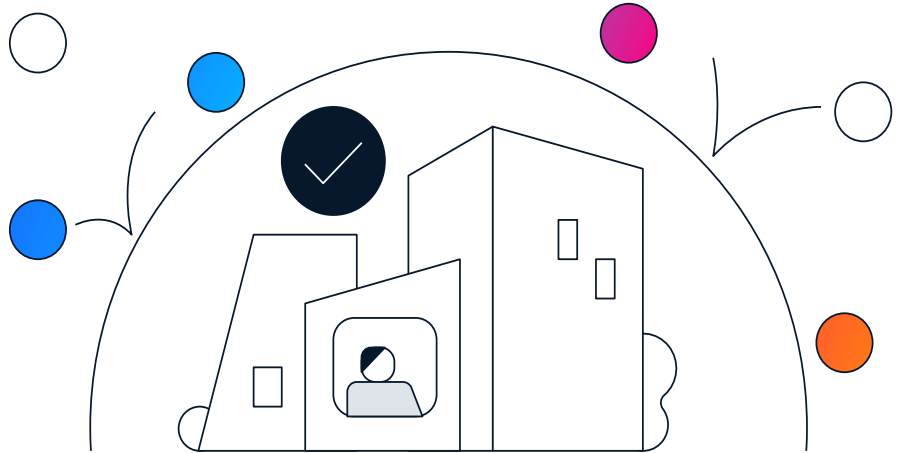
This includes tightening default configurations, enhancing logging and monitoring for richer security telemetry, and modernising device authentication with stronger protocols and encryption, all aimed at reducing the attack surface and helping customers anticipate and withstand tomorrow's threats. Together, these efforts demonstrate Cisco's commitment to not just reacting to emerging AI-era threats, but to getting ahead of them and helping our customers build a more resilient digital foundation.

Our early use of frontier AI models indicates that the legacy model of "one CVE per vulnerability" is reaching a breaking point. As automated discovery drives an exponential increase in identified bugs, treating every minor flaw as an individual disclosure record clogs the security ecosystem and actively delays software currency. Our north star is to empower customers with actionable intelligence, not overwhelming data. By pivoting toward a consolidated disclosure model through which severe vulnerabilities are prioritised and minor fixes are rolled into standard release cycles, we can accelerate patching decisions. This streamlined approach denies threat actors the detailed roadmaps they need to weaponise AI against our infrastructure.

To defend against modern threats, we must prioritise action over administration. The traditional approach of assigning individual CVEs to every minor issue creates a "vulnerability tax" that slows down upgrades and exhausts security teams. We believe the future of disclosure must focus on the outcome: guiding customers toward swift mitigation and upgrade motions. We need a strong CVE programme in the industry that can scale to this new level of security vulnerability discovery and disclosure.

These efforts demonstrate Cisco's commitment to not just reacting to emerging AI-era threats, but to getting ahead of them and helping our customers build a more resilient digital foundation.





Defending our enterprise

We are also applying these principles to our own enterprise environment. The recommendations outlined below are not theoretical; they reflect the same approach Cisco is taking internally to defend against AI-era threats. From accelerating patch cycles and eliminating end-of-life systems to deploying AI-assisted threat hunting and enforcing least privilege for AI agents, we are operationalising this guidance across our own infrastructure.

Our recommendations

To effectively respond to the accelerating capabilities enabled by advanced AI models, **organisations must adopt a balanced approach that reinforces foundational security practices while simultaneously modernising their defensive architecture.** While the threat landscape is evolving rapidly, many successful attacks still exploit well-known weaknesses. Strengthening core controls remains one of the most impactful actions security leaders can take.

Organisations should **prioritise foundational measures such as phishing-resistant authentication, strong identity verification, least privilege access (including AI agents), and zero-trust architectures.** Consistent patch management, comprehensive asset visibility, and disciplined configuration management are essential to reducing exploitable vulnerabilities. These controls form the baseline for resilience and are critical in limiting the scope and spread of both traditional and AI-era attacks. In many cases, improving execution on these fundamentals will deliver more immediate risk reduction than deploying new technologies alone.

At the same time, organisations must take an aggressive stance on eliminating structural risk. **Any devices or software that cannot be patched, upgraded, or supported must be systematically removed and replaced with modern platforms.** Modern systems incorporate advanced protections such as memory safety mechanisms and exploit mitigations that significantly increase the difficulty of weaponising vulnerabilities. Even when vulnerabilities exist, these protections slow attackers and reduce the likelihood of successful exploitation. Building environments that are flexible, continuously upgradable, and designed for rapid patching is now a critical requirement—particularly for internet-facing services, where very little time will be available between disclosure and mass exploitation.

But strengthening fundamentals and modernising infrastructure alone is insufficient. The speed of AI-driven attacks will compress the window between vulnerability discovery and exploitation to minutes or seconds. Traditional models based solely on detection and response are no longer adequate when used in isolation. **Defenders must evolve their operating model to match the speed, scale, and adaptability of AI-era threats.** This includes investing in machine-speed detection, automated triage and containment, and continuous monitoring of identity and data activity. This reduces reliance on manual intervention and enables faster, more consistent responses to high-confidence threats.

This evolution also requires a shift toward **embedded active defence**. Rather than relying exclusively on telemetry collection and post-event analysis, organisations should place protections directly within the workload, device, and traffic path, enabling security controls to act in real time. Examples include in-line enforcement mechanisms, runtime protections using technologies such as eBPF for low-level visibility and control, and independently updatable exploit shields that can respond to emerging threats without requiring full system upgrades. These capabilities must be designed to evolve rapidly, with the ability to update protections independently of major software or hardware refresh cycles.

Balance foundational controls with adaptive, real-time defence capabilities



Strengthen fundamentals

Phishing-resistant MFA, zero trust, least privilege (including AI agents), disciplined patch management, and full asset visibility.



Eliminate structural risk

Remove end-of-life systems. Replace with modern platforms featuring memory safety and exploit mitigations. Build for continuous upgradability.



Automate at machine speed

Invest in automated detection, triage, and containment. Manual-only response models can't match AI-driven attack velocity.



Embed active defence

Place protections in the workload, device, and traffic path—eBPF runtime controls, in-line enforcement, independently updatable exploit shields.



Harness AI for defence

Use AI for threat hunting, conformance testing, digital twins, and validation—compressing deployment cycles from months to days.

Organisations should also **harness AI capabilities for their own defence**. Constant internal threat hunting, aided by the same capable models adversaries utilize, will be a key capability for successful defenders. AI-driven conformance and acceptance testing can replace labour-intensive manual verification with high-velocity, automated intelligence, thereby generating complex test cases that cover edge cases often missed by human testers. In high-stakes environments, AI-driven digital twins can simulate production networks at scale, verifying that updates adhere to strict security protocols and performance benchmarks without risking the stability of live environments. Integrating AI into acceptance and validation phases significantly reduces the deployment bottleneck, **compressing the transition from code complete to field deployed from months to days**.

Ultimately, success in this new environment requires a dual focus: executing foundational controls with discipline while advancing toward adaptive, real-time, and embedded security capabilities. Organisations that aggressively reduce legacy risk, modernise their infrastructure, adopt an assume-breach mindset, and embrace active defence models will be best positioned to manage the speed and scale of AI-driven threats.

Conclusion

Change is coming. Defenders must take a sober look at the environment they are defending now and begin shaping that environment to survive in an AI-era adversary world. The wisdom of yesterday is still important, but it must be combined with modern, cutting-edge defensive capabilities, networks with exceptional visibility, and appropriate use of AI agents to assist humans in securing the environments they defend.