AI Defense



Artificial Intelligence: New Potential, New Risks

Artificial intelligence (AI) is rapidly reshaping the way organizations around the world operate. Business leaders are eager to adopt this transformative technology to drive greater productivity, enrich product capabilities, and gain a competitive edge—but this cannot come at the cost of safety and security.

AI security, meaning security for AI, is different than traditional cybersecurity because AI applications are fundamentally different from traditional software. This difference is especially stark with agents, which are AI systems that can plan and act autonomously. To protect enterprise systems that connect humans, agents, and sensitive business data, organizations need a secure approach that is purpose-built around AI.

© 2026 Cisco and/or its affiliates. All rights reserved.

AI Security Requires a New Approach

Since AI is non-deterministic, a single input (such as a prompt) can produce inconsistent outputs and actions. Because training data is part of the model itself, AI blurs the once-distinct lines between data and code. With agents in particular, AI is capable of autonomous action with little to no human supervision.

The same traits that make AI technology unique and powerful introduce new forms of organizational risk. At a high level, businesses embracing AI must contend with a number of security challenges:

- **Visibility gaps:** AI applications and their dependencies can exist in various locations including disparate cloud environments, codebases, and user endpoints. Without visibility, it becomes nearly impossible to govern and secure these assets.
- **Supply chain vulnerabilities:** AI applications frequently rely on components sourced from third parties including open-weight models, MCP servers, datasets, tools, and skills. These assets are susceptible to both inherent systemic vulnerabilities and malicious insertions.
- **Behavioral vulnerabilities:** It is difficult to predict how non-deterministic AI applications will perform in production settings, where they can exhibit harmful behaviors or fall victim to any number of attack techniques perpetrated by a bad actor.
- **Runtime attacks:** Adversaries will employ a broad arsenal of techniques to compromise enterprise AI applications and exfiltrate sensitive data, manipulate agents, maximize resource consumption, or execute some other malicious objective.

Traditional cybersecurity solutions are not designed for this new paradigm of AI risk. That's why we introduced Cisco AI Defense, a purpose-built solution for securing agents and AI applications from development through deployment. With capabilities that include AI supply chain scanning, adaptive AI red teaming, and customizable, industry-leading agentic AI guardrails, Cisco AI Defense enables organizations to embrace AI without compromising on security.

Cisco AI Defense

Our Framework

Cisco AI Defense takes an approach to AI security that mirrors the lifecycle of AI development. There are three parts:

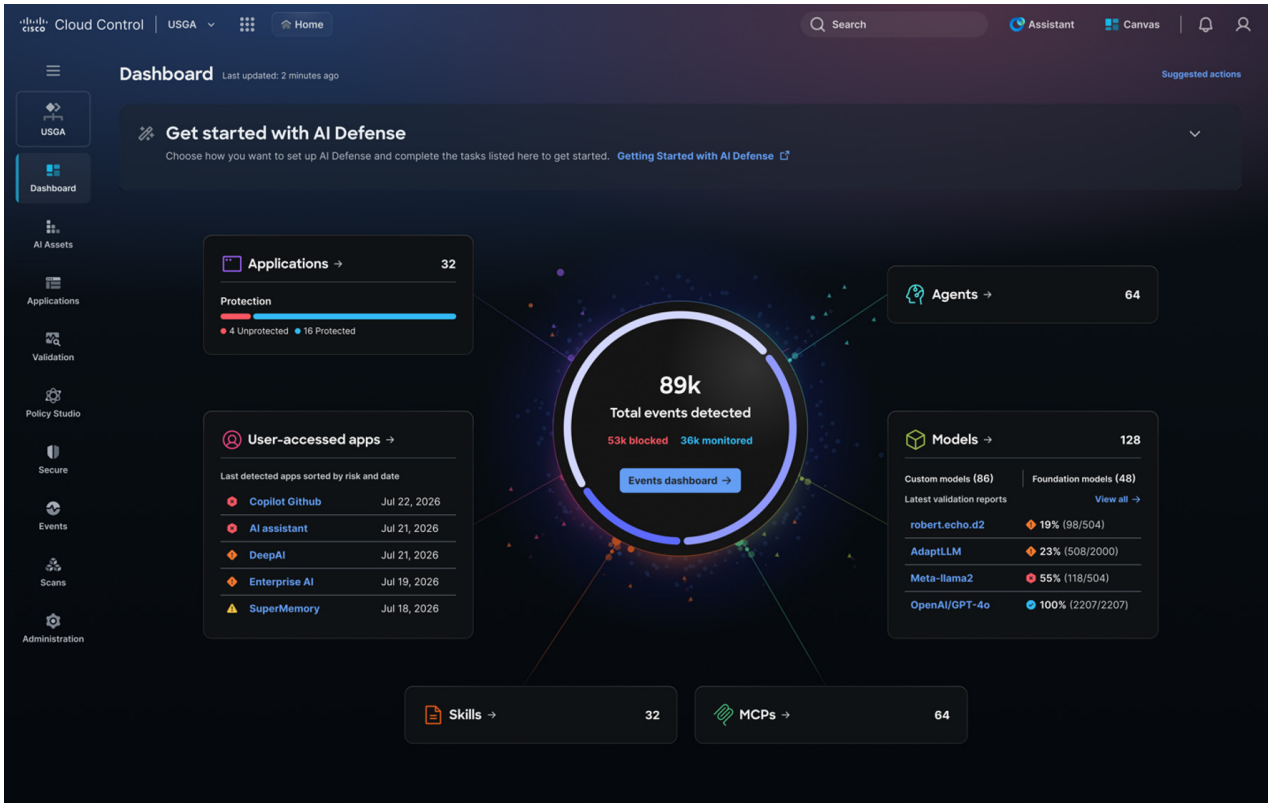


- **Discovery** of AI applications and all associated AI assets including agents, models, knowledge bases, MCP servers, tools, and skills. This visibility extends to every location where assets might reside—cloud environments, codebases, and user endpoints.
- **Detection** of vulnerabilities in the AI supply chain and in the behavior of the model or agent itself. This is achieved with continuous scanning of AI assets and algorithmic red team testing to measure behavioral risk.
- **Protection** of the agent or AI application against runtime threats such as prompt injection attacks, sensitive data exfiltration, or harmful content generation. AI Defense guardrails can be fully customized to fit the unique security requirements of any application, organization, or industry.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco AI Defense Capabilities

Cisco AI Defense delivers a full suite of capabilities for securing enterprise agents and AI applications from development through deployment. Each module in AI Defense supports the greater framework of Discovery, Detection, and Protection.



AI Inventory provides complete visibility and governance into the AI environment by discovering and inventorying AI assets wherever they reside. This visibility extends into cloud platforms like Amazon Bedrock, Google Cloud, or Microsoft Azure, code repositories like GitHub, and user endpoints. From AI Inventory, users can drill into additional details like agent composition details, asset metadata, model provenance, and security approval status.

AI Supply Chain Scanning helps ensure that agents and AI applications are built with secure components by identifying potential vulnerabilities in model files, MCP servers, tools, and skill files. Examples include model manipulation, tool poisoning, arbitrary code execution, and unauthorized system or network access.

AI Defense supply chain scans can analyze both an AI asset and its underlying intent; for example, a tool might claim to compose emails on the user's behalf while discreetly relaying every email to an external adversary. Even approved tools are susceptible to manipulation at a later time, which is why AI Defense offers routine scheduled scans for registered MCP servers.

AI Red Teaming simulates interactions with models, agents, and AI applications to expose complex behavioral vulnerabilities. These multimodal test for susceptibility to hundreds of adversarial techniques and content categories; every simulated attack is mapped to Cisco's comprehensive Integrated AI Security and Safety Framework, while also accounting for AI security standards from organizations like NIST, MITRE, and OWASP. AI Defense can detect threats in minutes that would have taken security analysts weeks to assess.

© 2026 Cisco and/or its affiliates. All rights reserved.

In addition to this broad-spectrum testing, AI Defense enables users to define custom objectives that reflect their unique security or safety concerns. The adaptive red team engine will interpret these objectives, then plan and execute a sophisticated multi-stage attack. Users can then see which attacks were successful and how easily they were achieved.

AI Runtime Protection protects agents and AI applications against threats in real time with bi-directional guardrails that filter user inputs and agent responses. Purpose-built guardrails also secure interactions between agents and connected tools.

AI Defense guardrails cover a broad spectrum of common AI risks including prompt injection, code detection, tool exploitation, sensitive data leakage, toxicity, and hate speech. For more specific threats unique to any given application, organization, or industry, users can leverage Policy Studio to build custom adaptive guardrails. The Policy Studio agent will converse with the user to refine the guardrail, improve the specificity and efficacy of its coverages, account for edge cases, and evaluate accuracy over time.

Altogether, the modules in Cisco AI Defense provide comprehensive security for agents and AI applications that span from the earliest stages of development all the way through runtime.

Why Cisco AI Defense

With a world-class AI team and decades of leadership in networking and security, Cisco has paved the way for rapid and secure enterprise AI adoption. several Here are a few reasons why Cisco AI Defense stands out in the market:

- **Industry-Leading Agent Security:** Cisco AI Defense out-of-the-box red team testing and guardrails cover the broad spectrum of common AI risks and are continuously updated with threat intelligence from Cisco Talos and our AI security research team. Plus, adaptive red teaming and custom guardrails in Policy Studio provide powerful personalization capabilities for customers to test for and protect against unique risks.
- **Support for Any AI Deployment:** Cisco AI Defense is AI infrastructure and application-agnostic, meaning it can deploy and run wherever your teams build today. By integrating with agent frameworks and platforms, AI Defense fits into existing AI development lifecycles without friction. It is validated on Amazon Web Services, Microsoft Azure, and Google Cloud; on neocloud GPU capacity; and on-premises in your own data center. With a data plane deployed inside your environment, inference traffic never leaves private infrastructure to meet data residency requirements without compromising on in-line enforcement.
- **Cisco Ecosystem Advantage:** AI security is done best at the network level, enforcing policies consistently across all AI traffic. Native integrations with first-party Cisco solutions and leading third-party solutions means AI Defense is seamless with your preferred agent security stack.

Learn more

Cisco remains at the forefront of AI security research and innovation as part of our broader commitment to championing a secure future for artificial intelligence. We regularly share open-source projects, AI security resources, original research, and product announcements across our AI blog and social channels.

Learn more about Cisco AI Defense at https://www.cisco.com/go/ai-defense.

© 2026 Cisco and/or its affiliates. All rights reserved. Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: www.cisco.com/go/trademarks. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)

C22-5114795-01 09/26