AI Defense



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

Contents

Secure your AI Transformation with Cisco AI Defense 3
The Cisco AI Defense solution 4
Cisco AI Defense advantages 4
AI Defense core components 5
AI Defense deployment options 9
Mapping to developing industry standards 9
Cisco Security for AI 10
Next steps 10

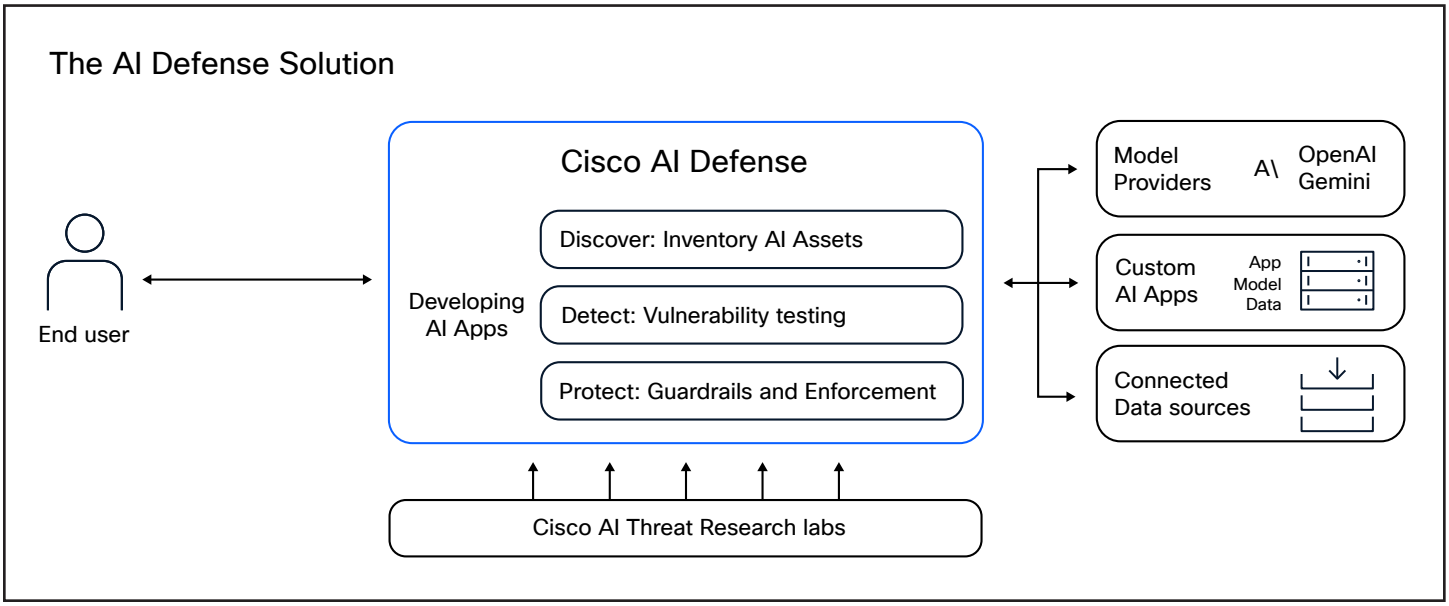
© 2026 Cisco and/or its affiliates. All rights reserved.

Secure your AI Transformation with Cisco AI Defense

Existing security controls are oblivious to AI infrastructure

As your organization produces AI-driven applications, you face significant challenges that existing security controls can't address. These include:

- **Visibility gaps:** Any number of AI models, agents, datasets, and other assets may exist across your distributed cloud environments.
- **AI supply chain exposure:** Compromised model files, repositories, and MCP servers can introduce hidden risks early in the development process.
- **AI model vulnerabilities:** Generative AI models are vulnerable to safety and security threats, despite built-in guardrails. Left unchecked, these underlying vulnerabilities can expose AI applications and agents to risks in production.
- **Emerging adversarial and agentic AI threats:** AI systems and autonomous agents introduce new classes of risk that existing controls do not address. These include prompt injection, jailbreaking, supply chain compromise uncovered through Detect scanning, tool misuse, misaligned objectives, privilege escalation, and deceptive or unsafe emergent behaviors.
- Protecting your organization from these threats requires a new approach.



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

The Cisco AI Defense solution

Cisco® AI Defense is a comprehensive, purpose-built AI security solution that mitigates the risks associated with the development and deployment of AI models, applications, and agents.

At its core, AI Defense empowers your security teams to discover AI assets across distributed cloud environments, detect vulnerabilities through supply chain scanning and algorithmic red teaming, and protect runtime applications with security, privacy, and safety guardrails updated with the latest threat intelligence.

AI Defense extends protection to agentic AI systems and the Model Context Protocol (MCP). This includes scanning MCP servers for compromised or malicious assets, enforcing runtime protections across MCP requests and responses, and detecting agent-specific threats such as memory poisoning, tool misuse, privilege escalation, intent hijacking, and deceptive agent behavior.

Cisco AI Defense advantages

Cisco AI Defense delivers three core advantages that help organizations secure AI applications at scale:

AI Model and Application Validation	AI Threat Intelligence and Standards Alignment	Network-Layer Enforcement Across Your AI Estate
Automatically test models and applications at AI scale with Cisco algorithmic red teaming. AI Defense evaluates AI asset performance against simulated attacks in 200+ threat subcategories, then generates model-specific guardrails to address issues such as prompt injection techniques, data privacy exposures, security weaknesses, and safety violations.	AI Defense is continuously updated with attack data from Cisco's AI threat research team and Talos, ensuring coverage for the latest adversarial techniques. Findings and protections map to leading frameworks, including MITRE ATLAS, OWASP Top 10 for LLMs, and NIST AI-RMF. Additionally, these findings align with Cisco's AI Safety and Security Framework, a security taxonomy for the agentic era.	Built on Cisco's industry-leading networking and security footprint, AI Defense provides broad visibility and control across cloud, VPC, and on-premises AI POD deployments, across AI application traffic and MCP-mediated interactions, including model requests, responses, and agent workflows, to prevent unsafe behavior, block harmful action chains, and stop privilege escalation before it impacts your environment.

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

AI Defense core components

AI Defense introduces a contiguous layer of AI security, privacy, and safety for real-time organization-wide AI risk management. This solution is based on four core components, each extended to support emerging agentic AI systems and the MCP:

- **AI Inventory:** Identify AI assets like models and agents in your cloud environments to discover and close gaps. AI Inventory automatically detects MCP servers and agent processes alongside models and applications. This includes mapping MCP-connected workflows and agent-to-tool interactions to uncover hidden or unsanctioned agentic behavior.
- **AI Supply Chain Risk Management:** Scan AI model files, repositories, and MCP servers for hidden risks before they are introduced into development or production. AI Defense identifies malicious code, poisoned data, unsafe tools, compromised components, and high-risk MCP assets using Cisco threat intelligence and proprietary scanning techniques. These findings generate per-asset risk scores that support risk-profile-based policies and preemptively block unsafe models or tools.
- **AI Red Teaming:** Leverage AI algorithmic red teaming to identify safety and security vulnerabilities quickly – tasks that traditionally require weeks of manual red teaming – and operate with the speed, adaptability, and comprehensiveness of an AI powered tool.
- **AI Runtime Protection:** Secure prompts and responses to AI applications and agents with safety, security, and privacy guardrails informed by the latest threat intelligence to protect against emerging AI threats, including prompt injection, DoS attacks, and sensitive data leaks. Runtime protection extends to MCP traffic, inspecting agent actions and tool calls in real time. This enables detection and blocking of unsafe agent behaviors such as unauthorized tool usage, harmful action chains, memory poisoning attempts, and other agent-driven risks.



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

AI Inventory

Features	Benefits
▪ **Automatic AI asset discovery:** Gain continuous visibility across AI-related cloud traffic (ingress, egress, and east-west) while discovering AI assets (e.g., models and agents) across your cloud environments, spanning Amazon Bedrock, Google Vertex, and Azure AI Foundry. ▪ **AI-traffic flow mapping:** Gain a complete view of your enterprise AI attack surface. ▪ **AI asset policy check:** Automatically track and prioritize assessed AI models, initiating scans and guardrails for unprotected assets. ▪ **See each model's AI security controls**: Quickly identify relevant security controls for further vulnerability testing and runtime protection. ▪ **MCP & agent discovery:** Automatically identify MCP servers, agent processes, and tool integrations operating in cloud environments to highlight hidden agentic workflows.	▪ Find models and AI assets no matter where they reside or when they arise. ▪ Eliminate the uncertainty and risk of rogue AI assets residing in your cloud with a single-pane-of-glass view that lets you automatically inventory your organization's AI security exposure. ▪ Identify new or unsanctioned assets and bring them into compliance with your security rules. ▪ Make informed decisions about security policies by mapping connections across data, models, and agents.

Supply Chain Risk Management

Features	Benefits
▪ **MCP server and tool scanning** – Analyze MCP servers and tools during supply chain scanning to identify vulnerabilities, assess their severity, and prevent unsafe or compromised components from entering development or production environments. ▪ **Model and repository scanning** – Scan AI model files and complete repositories to identify hidden risks such as malicious code, model backdoors, or embedded unsafe components, ensuring only trusted assets progress into development or production pipelines. ▪ **Risk scoring for supply chain assets** – Generate per-asset risk scores during supply chain evaluation to quantify exposure and support consistent, policy-driven gating decisions that block high-risk components before they are used.	▪ Detect early-stage agentic risks by verifying the safety and security of MCP servers and associated tools before they can introduce unexpected behavior or vulnerabilities. ▪ Prevent compromised or tampered models from entering development by identifying malicious code, poisoned data, or unsafe components during integrity scanning. ▪ Enable consistent, risk-based decision-making by applying per-asset risk scores that guide gating and enforcement policies across the AI lifecycle. ▪ Block unsafe or high-risk MCP tools and servers before use by enforcing allowlists and blocklists derived from supply chain scanning insights.

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

AI Model and Application Validation

Features	Benefits
▪ **Algorithmic red teaming:** Test the model, application and agent vulnerabilities at AI scale by automatically running 200+ attack techniques and threat subcategories using Cisco's proprietary AI: – 50+ attack procedures – 25 content categories – 10 languages – 3 modalities (text, image, audio) ▪ **Vulnerability reporting:** Reports include model, application and agent vulnerabilities. ▪ **Model-specific guardrails:** Generate guardrails tailored to specific vulnerabilities found in each model.	▪ Immediately identify hundreds of potential safety and security risks across both models and agents in real time. ▪ Find susceptibility to malicious actions, such as prompt injections, data poisoning, or unintentional outcomes. ▪ Track risks, vulnerabilities, and threats across your entire AI-enabled application stack, including models and agents. ▪ Simplify compliance with automated reporting that translates test results into an easy-to-read report mapped to industry and regulatory standards.

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

AI Runtime Protection

Features	Benefits
▪ **Safety, security, and privacy controls:** AI Runtime inspects all prompts and responses to block violations. ▪ **Security:** Protect against AI attacks and threats like prompt injections, denial of service, malicious URLs, and more. ▪ **Privacy:** Prevent sensitive information leakage, including Personally Identifiable Information (PII), Payment Card Industry Data Security Standard (PCI), Protected Health Information (PHI), source code, model information, and more. ▪ **Safety:** Block content that can be toxic or cause harm around financial, societal, reputational, and user categories. ▪ **MCP runtime enforcement:** Monitor and enforce safe behavior across MCP requests and responses connecting LLMs, agents, and tools. Runtime policies now enforce tool allowlists and blocklists, automatically restricting MCP server or tool usage based on assigned risk profiles and severity scoring from MCP validation workflows. ▪ **Agent behavior protection:** Detect and block unsafe agent actions, including unauthorized tool usage and privilege escalation. ▪ **Events dashboard:** Achieve visibility into policy guardrail violations with information around event types, prompts, responses, and more. ▪ **Multiple enforcement points:** AI Defense guardrails cover multiple traffic types, including API calls and outbound and internal communications between services and AI applications.	▪ Ensure safe outputs by blocking sensitive information and harmful content. ▪ Ensure that AI-enabled applications remain compliant. ▪ Security teams can tailor rules to organizational standards and tolerances. ▪ Integrate with AI security operations and broader SecOps through security console integrations, such as Splunk. ▪ Prevent harmful or unintended agent actions by enforcing policies on tool invocation, privilege levels, and action chains. ▪ Gain complete visibility and control over agentic workflows, including MCP-mediated communication.

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

AI Defense deployment options

In all VPC and AI POD deployments, AI Defense ensures that application, model, and agent data remain fully within the customer's environment. Only metadata required for management-plane operations is sent to Cisco, while all sensitive inference traffic, MCP interactions, agent workflows, and runtime policy enforcement stay entirely in the customer's cloud or on-premises infrastructure.

Mapping to developing industry standards

Keep on top of the AI revolution: AI Defense helps your teams stay ahead by aligning with evolving standards like NIST AI Risk Management Framework (AI-RMF), MITRE ATLAS, and OWASP Top 10 for LLM, and OWASP Top 10 for Agentic AI. This standards alignment extends across model, agent, and MCP workflows, including supply chain risk categories for model files, datasets, tools, and MCP servers. These protections now extend to emerging guidance for agentic AI systems and protocols, such as the MCP, ensuring that both models and agents meet evolving security expectations.

The Cisco AI Defense research team actively contributes to these standards and participates in industry working groups, helping your organization stay compliant with new regulations and best practices as they develop.



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

Cisco Security for AI

Cisco is building on decades of leadership in networking and cybersecurity to pave the way for rapid AI innovation and resilient AI security. We cover areas across:

- **AI apps built by the enterprise:** Cisco AI Defense protects against the safety and security risks introduced by the development and deployment of AI applications, models, agents, and MCP servers, providing end-to-end protection from development to runtime.
- **Third-party GenAI apps or Shadow AI:** Cisco safeguards organizations from the security risks of third-party AI applications with Cisco Secure Access, protecting against threats and sensitive data loss while restricting employee access to unsanctioned tools.
- **AI supply chain and risk management:** Cisco secures the AI supply chain across open-source and enterprise environments. Through partnerships such as HuggingFace, Cisco scans public model files using Talos threat intelligence and ClamAV to identify malicious code and embedded risks. AI Defense extends these protections into private customer repositories by scanning models, datasets, and MCP assets before they enter development pipelines, and enforcing security across agent workflows at runtime.

Next steps

Learn more at 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)