AI Defense

Contents

Secure Your AI Transformation with Cisco AI Defense....3

The Cisco AI Defense Solution....4

Cisco AI Defense Advantages....4

AI Defense Core Features and Benefits....5

AI Defense Deployment Options....7

Mapping to Developing Industry Standards....8

Cisco Security for AI....8

Next steps....8

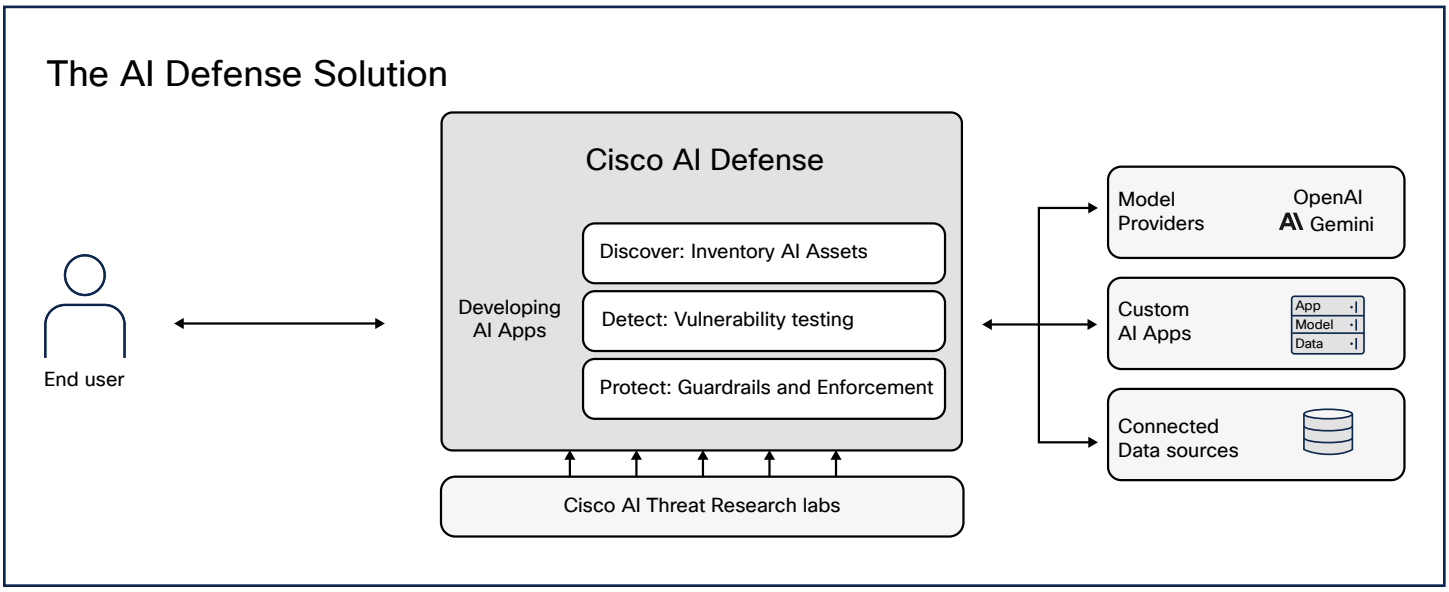
Secure Your AI Transformation with Cisco AI Defense

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:** AI infrastructure is multi-cloud, multi-model, and multi-modal, leaving significant gaps in coverage.
- **AI model vulnerabilities:** AI models are non-deterministic, requiring continuous assessment and validation.
- **Emerging adversarial AI threats:** AI systems are vulnerable to novel attack vectors targeting runtime applications (e.g., prompt injections and jailbreaking).

Protecting your organization from these threats requires a new approach.



The Cisco AI Defense Solution

Cisco® AI Defense mitigates the risks associated with AI development, deployment, and usage by embedding industry-leading AI and cybersecurity technology into the Cisco Security Cloud.

At its core, AI Defense empowers your security teams to detect vulnerabilities, implement real-time guardrails, and seamlessly integrate across the enterprise AI attack surface.

Cisco AI Defense Advantages

Assess vulnerabilities and risk with AI Model and App Validation	Stay ahead of adversarial threats with AI threat intelligence	Enforce security at the network layer
AI algorithmic red teaming delivers an automated assessment of security and safety vulnerabilities with a report on recommended guardrail enforcement.	Cisco's AI threat research team continually updates AI Defense with the latest attack data as well as aligning with MITRE Adversarial Threat Landscape for AI Systems (ATLAS), Open Web Application Security Project (OWASP), National Institute of Standards and Technology (NIST), and other frameworks.	Leverage Cisco networking expertise with broad visibility and enforcement points across your AI attack surface area.

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 three core components:

- **AI Cloud Visibility:** Identify AI assets like models and agents in your cloud environments to discover and close gaps.
- **AI Model and Application Validation:** Leverage AI algorithmic red teaming to identify safety and security vulnerabilities quickly – tasks that traditionally require weeks of manual red teaming.
- **AI Runtime Protection:** Secure AI applications with safety, security, and privacy guardrails that automatically adapt to emerging AI threats such as prompt injections, DoS attacks, jailbreaking, and sensitive data leaks.

AI Cloud Visibility

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, such as Amazon Bedrock. ▪ **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.	▪ 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.

AI Model and Application Validation

Features	Benefits

• **Algorithmic red teaming** – Test the model vulnerabilities by automatically running 200+ attack techniques and threat categories using Cisco's proprietary AI:
  – 45+ prompt injection techniques
  – 30+ data privacy categories
  – 20+ security targets
  – 50+ safety categories
• **Vulnerability reporting** – Reports include prompt injection attack techniques, data security, privacy, and safety vulnerabilities.
• **Model-specific guardrails** – Generate guardrails tailored to specific vulnerabilities found in each model.
	• Immediately identify hundreds of potential safety and security risks 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 (e.g., models, data, users, apps) via an API or Software Development Kit (SDK.)
	• Get AI security risk reports across to assess compliance and the status of AI applications.

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.
	• 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.

AI Runtime Protection	
Features	**Benefits**
• **Violations dashboard** – Achieve visibility into violations with information around violation 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.	

AI Defense Deployment Options

Components	Validation Essentials	Runtime Essentials	Advantage
AI Cloud Visibility	✓	✓	✓
AI Model and Application Validation	✓	-	✓
AI Runtime Protection	-	✓	✓

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. Even better, the AI Defense team actively contributes to these frameworks, preparing your organization for emerging regulatory and industry requirements.

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.
- **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 protects against malicious AI model files (e.g., scanning for malicious code, software license compliance, and geopolitical origin risks) entering the enterprise through network-based enforcement with Cisco Secure Access, Cisco Secure Endpoint, and Cisco Email Threat Defense.

Next steps

Learn more at https://www.cisco.com/go/ai-defense.