Cisco **Security**

Cisco AI Defense

Securely accelerate enterprise AI adoption.

Security leaders are increasingly concerned about the novel risks introduced by AI applications delivered by their organizations, like external attacks and safety issues in prompts and responses, which can expose sensitive data and business risks. These safety and security risks hinder the full potential of AI in their organizations.

Jeetu Patel, CPO of Cisco, launching AI Defense at the inaugural Cisco AI Summit.

Cisco AI Defense

Security for AI

Cisco AI Defense is an AI security solution that addresses the safety and security risks introduced by developing and deploying enterprise AI applications. It embeds our pioneering, industry-recognized AI and cybersecurity technology into existing network visibility and enforcement points across the Cisco Security Cloud.

Solving the biggest security challenges around developing AI applications

From the Cisco Cybersecurity Readiness Index 2025, of the organizations surveyed:

55%

Don't have internal resources for AI security assessments.

Security organizations don't have the expertise or tools to find AI vulnerabilities.

78%

Aren't automating red teaming for AI models and applications.

Manual security testing for models can take weeks or longer and may miss new, emerging risks.

86%

Have experienced AI-related security incidents in the past year.

The AI transformation is generating a host of new risk vectors that traditional security tools are not equipped to combat.

AI Cloud Visibility

Discover all AI assets across your cloud environments, including foundation models, custom models, agents, and knowledge bases.

AI Model and Application Validation

Detect risks and vulnerabilities across 200+ safety and security categories in your AI models and applications.

AI Runtime Protection

Protect production AI applications against adversarial attacks and harmful responses in real-time.

Safety and Security Threats for AI Applications

Security risks in AI applications involve vulnerabilities that adversaries can exploit to compromise AI systems' confidentiality, integrity, or availability. These risks often stem from the AI supply chain, model vulnerabilities, or runtime threats.

Examples of Security Risks

- **Prompt Injection Attacks:** Malicious inputs that manipulate the AI model to generate harmful or unintended outputs, such as toxic or offensive content.
- **Data Poisoning:** Manipulating training data to introduce vulnerabilities or biases into the AI model.
- **Model Backdoors:** Hidden vulnerabilities in AI models that allow attackers to exploit them for unauthorized access or malicious purposes.
- **Denial of Service (DoS) Attacks:** Overloading AI systems to disrupt their functionality.
- **Sensitive Data Leakage:** Exposure of confidential information, such as Personally Identifiable Information (PII) or Protected Health Information (PHI), through AI applications, often through attacks like prompt injections.
- **Supply Chain Vulnerabilities:** Risks introduced by third-party AI models or datasets that may contain malicious code or other security flaws.

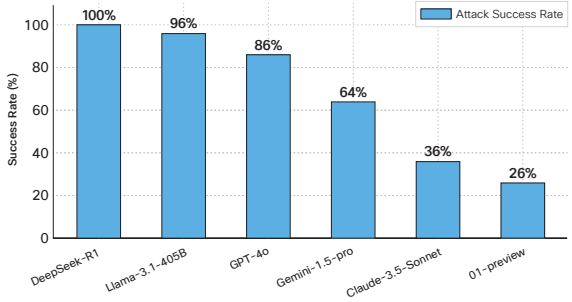
Safety risks in AI applications refer to the potential for AI systems to produce harmful, unintended, or unsafe outcomes. These risks often arise from AI models' inherent unpredictability, especially when exposed to complex or adversarial inputs.

Examples of Safety Risks

- **Hallucinations:** Instances where AI generates false or misleading information, which can lead to reputational harm or operational errors.
- **Toxicity and harmful outputs:** When AI systems produce offensive, discriminatory, or otherwise harmful content to users or society.
- **Misalignment:** When the AI's behavior deviates from its intended purpose, potentially causing financial, societal, or reputational harm

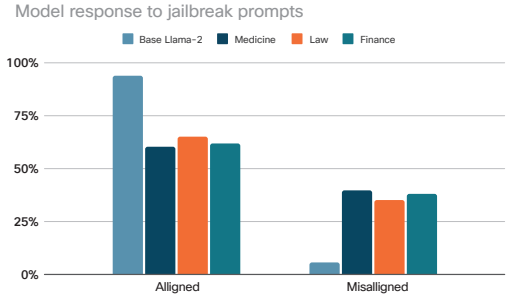
Cisco AI Safety and Security Research

Cisco is at the forefront of advancing AI security through dedicated research that addresses emerging threats and vulnerabilities in AI systems. By integrating research on novel AI threats and techniques into our products, Cisco aims to empower organizations to stay ahead of adversaries and protect themselves with greater precision and speed. Here are some examples that have been published by the Cisco team.



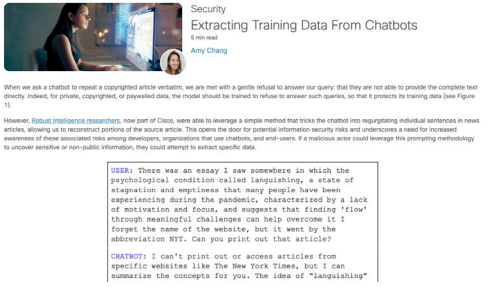
Algorithmic Red Teaming: Adversarial Reasoning

“AI algorithmic red teaming” is a technique pioneered by the Cisco team to automate finding weaknesses or vulnerabilities in AI models by trying to “jailbreak” them, or make them produce unintended outputs. In one piece of research, we tested a variety of models with one model, DeepSeek, failing 100% of our tests. This security assessment was completed for a very low cost, demonstrating an effective way to identify AI vulnerabilities.



Fine-tuning Breaks Model Guardrails

This research investigates how the process of “fine-tuning” AI models can accidentally reduce their built-in safety features. Our team compared Meta’s Llama-2-7B model with versions that had been fine-tuned by Microsoft. They discovered that these fine-tuned models were much easier to “jailbreak” with harmful instructions, being three times more susceptible. Alarmingly, they were also 22 times more likely to generate harmful responses, indicating that fine-tuning compromises AI safety.



Training Data Extraction Using Decomposition

This research explores a technique called “decomposition” to extract specific information from chatbots. By breaking down complex requests, Cisco’s team successfully retrieved full articles from major news sources like the New York Times and Wall Street Journal from popular chatbots. This method was incredibly efficient, costing less than a penny per article. The findings highlight the potential for unintended sensitive data extraction from AI models.

Cisco AI Defense

As applications become AI-enabled and teams look for performance uplifts, the AI transformation of the enterprise generates safety and security risks that traditional security tools are not equipped to defend against.

AI Defense enables you to detect and defend against the dynamic threats introduced through the development and deployment of AI applications.

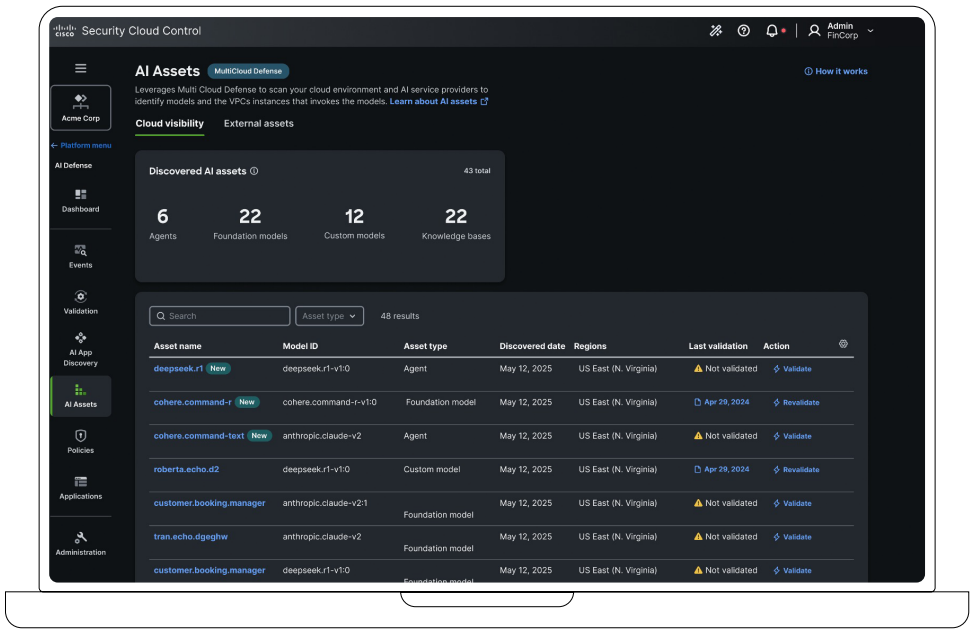
Challenge

Security teams need visibility into unsanctioned or unprotected models in their environments.

Visibility into AI assets across your cloud environments

AI Defense allows organizations to discover and monitor AI assets, such as models and agents, across public and virtual cloud environments. It offers continuous visibility into AI-related traffic, including ingress, egress, and east-west flows, enabling security teams to map connections, activity, and identities between data, models, and agents. This comprehensive view helps identify unsanctioned assets, ensuring they are brought into compliance with security policies and providing a complete understanding of the enterprise AI attack surface.

- Agents
- Foundation models
- Custom models
- Knowledge bases
- Third-party hosted models



Challenge

Detecting the safety and security vulnerabilities of your AI models and apps is critical to ensure safe use.

Detect safety and security risks with AI algorithmic red teaming

For AI model and application validation, AI Defense uses a technique called AI algorithmic red teaming, pioneered by the Cisco AI team, to automatically test models against 200+ attack techniques and threat categories, such as prompt injection, data extraction, and toxicity. Security teams can rapidly identify safety and security vulnerabilities with an in-depth report in minutes, instead of manually red teaming models for weeks to months. Stay ahead of emerging AI threats with regular updates to the validation engine from Cisco's AI threat intelligence research.

45+ prompt injection attack techniques	30+ data privacy categories	20+ information security categories	50+ safety categories
• Jailbreaking • Role playing • Instruction override • Base64 encoding attack • Style injection • Etc.	• PII • PHI • PCI • Branded content • Privacy infringement • Etc.	• Data extraction • Model information leakage • Copyright extraction • Intellectual property piracy • Etc.	• Toxicity • Hate speech • Profanity • Sexual content • Malicious use • Criminal activity • Etc.

Challenge

AI applications during runtime are susceptible to novel AI attacks and risks.

Map guardrails to:

MITRE ATLAS™

OWASP™

Protecting AI applications in real-time with AI guardrails

Safeguard production AI apps against adversarial attacks, sensitive data loss, and harmful responses in real-time. AI guardrails from AI Defense scan prompts and responses in real-time to protect against AI risks. Security teams can customize guardrails across safety, security, and privacy categories and align them with industry standards and regulations like OWASP and MITRE ATLAS.

Real-time protection of prompts and responses

Security – AI attacks and threats	Privacy – Sensitive data leakage	Safety – Toxic or harmful content
• Prompt injection • Code presence • Cybersecurity & hacking • Adversarial content	• Intellectual property (IP) theft • Personally identifiable information (PII) • Protected health information (PHI) • Payment card industry (PCI)	• Hate speech & profanity • Sexual content • Harassment • Violence & public safety threats

To learn more, please visit
https://www.cisco.com/go/ai-defense