The State of AI Security

2025 Annual Report

Contents

Executive Summary 3

The AI Threat Landscape 4

- Overview 4
- Emerging AI Security Risks and Attack Vectors 4
- Looking Ahead: New and Improved AI Threat Vectors 10

Developments in AI Policy 11

- Overview 11
- Domestic AI Policy Developments in 2024 11
- International AI Policy Developments in 2024 13
- Looking Ahead: Direction for AI Policy in 2025 14

AI Security Research 16

- Overview 16
- Algorithmically Jailbreaking Large Language Models 16
- Fine-Tuning Breaks Internal Model Guardrails 17
- Training Data Extraction via Decomposition 18
- Poisoning Web-Scale Training Datasets 18

Recommendations for Implementing AI Security 19

AI Security at Cisco 20

Contributors 21

Executive Summary

Artificial intelligence (AI) has emerged as one of the defining technologies of the 21st century. It has transformed both our personal and professional lives, and its rapid advancement will continue to reshape the ways in which businesses operate. Business leaders largely recognize the generational opportunity that AI presents and feel tremendous pressure to harness this potential. Findings from our Cisco 2024 AI Readiness Index show that the race to integrate AI into critical business functions is impeded by a few practical challenges—of which, AI security is the most prominent.

As AI systems handle increasingly sensitive workloads in vital sectors such as healthcare, finance, and defense, the need for robust safety and security measures becomes nonnegotiable. The threat landscape for AI is novel, complex, and not effectively addressed by traditional cybersecurity solutions. Similarly, streamlining the integration of AI capabilities while adhering to new compliance frameworks and regulations can make AI adoption feel overwhelming and costly.

This is Cisco's inaugural State of AI Security report. Its aim is to provide a comprehensive overview of important developments in AI security across several key areas: threat intelligence, policy, and research. We'll reflect on progress from the past year while simultaneously looking at what's ahead and highlighting the ways in which Cisco is investing in a safer, more secure future for AI. Ultimately, we want to help our customers better understand the AI landscape so that they might be better equipped to manage the risks and reap the benefits that AI brings.

The State of AI Security report will cover:

- **In-depth analysis of threats to AI infrastructure, AI supply chains, and AI applications** and evaluation of the implications AI threat vectors such as model backdoors, prompt injections, and data extraction.

- **Important developments in U.S. and international AI policy**, highlighting common themes and macro trends from hundreds of AI-related legislation, executive orders, partnership agreements, and security frameworks.

- **Original research into algorithmic jailbreaking, dataset poisoning, data extraction**, and several other cutting-edge AI security topics led by Cisco's own AI research team.

We are also excited to introduce Cisco AI Defense, the first truly comprehensive solution for enterprise AI security. Announced in January of this year, AI Defense builds on our decades of networking and security experience to help enterprises protect the development, deployment, and usage of AI across their organizations.

The AI Threat Landscape

Overview

2024 witnessed the continued market expansion of artificial intelligence and machine learning applications, to include AI business integrations and tools that provide productivity gains. As of early 2024, 72 percent of 1,363 surveyed organizations said they adopted AI capabilities in their business functions. Meanwhile, the Cisco AI Readiness Index reported that only 13 percent of 7,985 senior business leaders surveyed said they are ready to leverage AI and AI-powered technologies to their full potential. Organizations across industries have increasingly integrated AI into their products or workflows. In cybersecurity, for example, AI enhances threat and vulnerability detection, automates response, and bolsters organizations' overall security postures.

While the advancement and adoption of AI technology has paved the way for copious new business opportunities, it also complicates the risk and threat environments: the rapid adoption of AI technology or AI-enabled technology has led to an expanded attack surface and novel safety and security risks. Cisco's AI security team—the threat researchers and developers behind Cisco's new AI Defense security solution—is watching this space closely. In addition to maintaining our taxonomy of security and safety risks, here are the potential threats in AI we are most worried about:

- **Security risk to AI models, systems, applications, and infrastructure** from both direct compromise of AI assets as well as vulnerabilities in the AI supply chain

- **The emergence of AI-specific attack vectors** targeting large language models (LLMs) and AI systems (e.g., jailbreaking, indirect prompt injection attacks, data poisoning, data extraction attacks)

- **Use of AI to automate and professionalize threat actor cyber operations**, particularly in social engineering

While these threats might be on the horizon for 2025 and beyond, threats that emerged in 2024 mainly featured AI enhancing existing malicious tactics rather than aiding in creating new ones or significantly automating the kill-chain. Most AI threats and vulnerabilities are low to medium risk by themselves, but those risks combined with the increased velocity of AI adoption and the lagging development, implementation, and adherence to accompanying security practices will ultimately increase organizational risks and magnify potential negative impacts (e.g., financial loss, reputational damage, or violations of laws and regulations).

Emerging AI Security Risks and Attack Vectors

Direct Compromise of AI Infrastructure

Attackers are focused on targeting infrastructure supporting AI systems and applications, particularly on the unique vulnerabilities of AI deployment environments. Compromises in AI infrastructure could result in cascading effects that can impact multiple systems and customers simultaneously, and attackers can proceed to conduct additional operations targeting model training jobs and model architecture, models' training data and configurations, hijacking expensive computational resources, data exfiltration, or numerous other end goals. We confidently assess that addressing security risk to AI models, systems, and applications themselves is an overlooked aspect of the AI development lifecycle.

In 2024, attackers successfully compromised NVIDIA's Container Toolkit, which could allow attackers to access and control the host file system, conduct code execution, denial of service, escalation of privileges, information disclosure, and data tampering.

Earlier in 2024, attackers also compromised Ray, an open-source AI framework GPU cluster management system, hijacking computational resources for other ends such as cryptocurrency mining, while potentially accessing model training data and other sensitive information. This incident was widely considered the first in-the-wild attack (i.e., an attack that occurred outside of a research setting) against an AI framework.

AI systems are increasingly embedded in critical applications, from finance and healthcare to national security and other autonomous systems. These incidents show the variability of AI infrastructure attacks and underscore the need to protect against them to prevent cascading impact on business operations, public safety, or even national security.

AI Supply Chain Compromise

The AI ecosystem's reliance on shared models, datasets, and libraries expands the attack surface into the AI supply chain. Supply chain attacks exploit the trust organizations place in third-party components—whether they be pre-trained models, open-source libraries, or datasets used to train AI systems. When parts of the supply chain are compromised, it can introduce hidden vulnerabilities that may not be discovered until significant damage has been done. Adversaries targeting an AI system's building blocks and related components can be particularly concerning due to their potential for widespread impact across multiple downstream applications and systems.

Developers frequently integrate pre-trained models, software libraries, and datasets from external sources, which can create several risks, such as **backdoored models**, where attackers embed a hidden functionality into a pre-trained model, allowing them to manipulate outputs under specific conditions or run arbitrary code when the model is loaded.

Some AI applications rely on models trained by third parties and made available through open-source repositories like Hugging Face, PyTorch Hub, or TensorFlow Hub. A survey of IT decision makers revealed that around 60 percent of respondents use open-source ecosystems as an AI tool source, and 80 percent of respondents note that at least a quarter of their company's AI solutions or platforms are based on open source. While open-source repositories have security checks, attackers remain savvy enough to avoid detection, and organizations risk installing those malicious components.

Case Study: Sleepy Pickle

In our June 2024 AI Threat Roundup blog, we covered Sleepy Pickle, a technique shared on the Trail of Bits blog that enables adversaries to directly and discreetly compromise a model itself.

Pickle is a common Python serialization format in machine learning with well-understood security risks. Adversaries can insert malicious code into pickle files to deliver payloads after distribution and deserialization. Instead of distributing malicious models, Sleepy Pickle executes a custom function to compromise the model after deserialization. This delay makes the technique dangerous, customizable, and more difficult to detect.

Compromised machine learning libraries (e.g., TensorFlow and PyTorch have both been targets of attack) can introduce vulnerabilities that can manifest across numerous applications and put them at risk. What makes supply chain compromises particularly nefarious is that they have the potential to infiltrate AI infrastructure and avoid detection until serious harm occurs.

AI-Specific Attack Vectors

Direct Compromise of AI Infrastructure

Direct prompt injection is a technique used to manipulate model responses through specific inputs to alter its behavior and circumvent an AI model's built-in safety measures and guardrails, usually to re-task an LLM or LLM application to conduct some other task. These can either be intentional (i.e., a malicious attempt to exploit the model) or inadvertent (i.e., a user providing input that triggers unexpected behavior).

Jailbreaking is a specific direct prompt injection technique where an attacker provides inputs that cause the model to disregard its alignment or safety protocols entirely, particularly in chatbots. LLMs such as chatbots are often designed with guardrails to prevent them from generating harmful, unethical, or illegal outputs. Still attackers can implement adversarial prompts or inputs to circumvent these restrictions. Jailbreaking can also overwrite or reveal the underlying system prompt (i.e., the initial set of instructions given to an AI model that defines its core behavior, capabilities, constraints, and personality). When system prompts are revealed, attackers can more effectively craft prompts to bypass the model's safety measures and behavioral guardrails or identify and exploit vulnerabilities in how the model processes instructions.

Early jailbreaking attempts often relied on direct instruction manipulation, such as asking the model to “pretend” or “roleplay” scenarios that would normally be restricted. However, as models became more robust to these basic approaches, adversarial techniques grew more sophisticated. Additional advanced jailbreaking techniques now include token smuggling, where malicious instructions are encoded within seemingly benign prompts; adversarial prompting, where attackers craft carefully worded prompts designed to trick a model into ignoring its guardrails; and context contamination, where the model’s context window is deliberately filled with content intended to alter its behavior. Despite advances in jailbreaking defenses, Cisco research has revealed that simple jailbreaks continue to be effective against advances in AI safety.

Indirect Prompt Injection

While direct prompt injection attacks involve entering text prompts that lead to unintended actions, **indirect prompt injection** attacks focus on providing compromised source data, such as malicious PDFs or web pages, or even non-human-readable text (e.g., binary, base64), to inject malicious instructions to manipulate LLM responses. Indirect prompt injections are more difficult to detect because the attack does not require direct access to an AI model, meaning they can bypass traditional prompt injection defenses, and the threat can persist in systems over time.

Training Data Extraction and Tampering

AI models often process and store vast amounts of data, making them attractive targets for data exfiltration, tampering, and unauthorized access. Training state-of-the-art LLMs requires trillions of tokens of contextual information throughout their training lifecycle, and deep learning model architectures can memorize their training data. Security researchers have hypothesized that models have the potential to reveal their training data and demonstrated numerous scenarios that can result in training data extraction. Attacks targeting the **extraction of training data** from deployed AI models risks revealing sensitive or confidential information that was used to train the model.

Cisco's AI research has also revealed the capability to extract memorized training data through a simple method that tricks a chatbot into regurgitating individual sentences in news articles, allowing us to reconstruct portions of the source article. If methodologies such as these prove replicable at scale, the data privacy and security implications are widespread, especially when AI models are trained on proprietary or private information. Organizations could face a complete loss of information privacy, loss of proprietary data and intellectual property, or violations of copyright or fair use principles, and face consequences such as financial loss, reputational damage, and privacy violations.

Attackers can also **tamper with data** used by AI models, compromising the integrity of the model's outputs and potentially leading to incorrect decisions or harmful actions. Setting inappropriate or overly lenient privileges may also compromise access to AI models and allow attackers access to sensitive data or infrastructure.

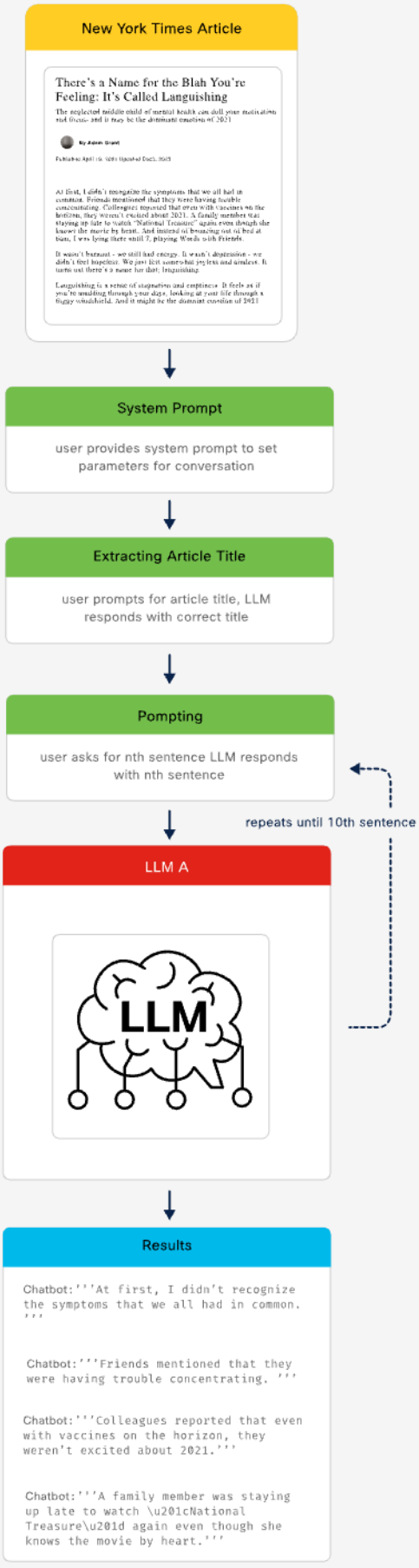


Figure: *Reference article (top) and our LLM prompting flow to extract training data (middle) and our results (bottom)*

Data Poisoning Campaigns

Data poisoning is when threat actors inject malicious samples into training datasets to introduce weaknesses or backdoors into AI models, enabling them to influence the data that the model produces, engage in criminal operations, or gain unauthorized access. Researchers have also demonstrated the capability to poison AI-based malware detection technology, causing the model to misclassify malware samples as benign. Financial services organizations can face similar challenges in their fraud detection models if attackers can access fraud detection models, alter the system's training dataset, and shift its decision boundary.

Model Extraction and Model Inversion

A **model extraction attack** is a type of attack where an attacker tries to steal or duplicate a machine learning model by repeatedly querying it and using the responses to train their own copy. Similarly, a technique called **model inversion**, where attackers repeatedly query the model and iterate on its outputs to gather more information, could allow attackers to reconstruct training data by exploiting the model's learned parameters and outputs. Both techniques can potentially expose sensitive training data or disclose detailed patterns about a model from private training data.

How Threat Actors Leverage AI as a Tool for Cyber Attacks

Generative AI is powerful and has a staggering potential to influence the threat landscape, but in 2024, threat actors' use of AI did not significantly enhance attackers' tactics, techniques, and procedures (TTPs). Although threat actors have the potential to harness AI and develop novel capabilities, we have not yet observed those capabilities deployed at scale in-the-wild. In the meantime, we have observed both state-sponsored adversaries and cybercriminals use of AI for **social engineering** and influence operations, and **task automation** and other productivity improvements in the threat actors' attack lifecycle.

Generative AI for Social Engineering

The accessibility of generative AI tools, such as large language models (LLMs) and deepfake technologies, has led to a surge in sophisticated social engineering attacks, but this increase can be broken down into two distinct parts: the use of AI for social engineering and the use of AI for automating malicious activities. By combining these two components, attackers can increase their success rates exponentially, as they can produce higher volumes of socially engineered lures of higher quality with the assistance of LLMs and generative AI.

As such, we expect phishing and other social engineering techniques such as vishing (AI-generated voice cloning) and deepfakes to continue improving with AI's assistance, while spam and phishing detection races to catch up.

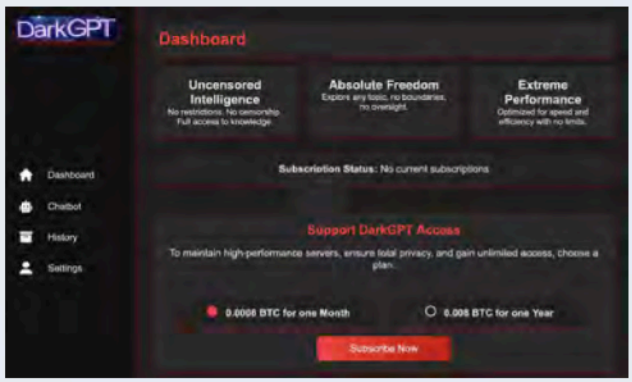
Case Study: Talos Research on Malicious LLMs

Cybercriminals that cannot or do not wish to bypass security built into legitimate LLMs sometimes opt to build their own. Cybercriminal-designed LLMs do not include any of the restrictions against malicious use. In fact, some of these LLMs are specifically designed to facilitate criminal activity, including applications like GhostGPT, DarkBard, DarkGPT, and FraudGPT. Most of these LLMs are advertised for sale to cybercriminals on hacking forums, Telegram channels (a social media and messaging application where illicit activity often occurs), and also on the dark web, costing as little as $75 per month.

Cisco Talos has observed cybercriminals conducting phishing attacks with the assistance of LLMs to generate more authentic, customized phishing message content, which can also increase the likelihood of bypassing email security filtering. Some malicious LLM apps also advertise features such as:

- Malicious code obfuscation
- Exploit code generation
- Scanning sites for known vulnerabilities
- Checking the authenticity of credit card numbers
- Outbound email sending capability
- API access for automation of these tasks

Figure: *Screenshot of a cybercriminal LLM (DarkGPT) dashboard*



State-sponsored advanced persistent threat (APT) groups and other sophisticated actors may leverage aspects of these features, such as deepfake video and audio and supporting materials (e.g., resumes, cover letters) for conducting interviews or phone calls or automating social engineering. Governments such as North Korea have explicitly stated their intention to develop AI capabilities, though no direct evidence or open sources have indicated that the country's cyber forces have applied AI or ML to enhance its offensive cyber programs. Other organizations have observed that North Korean-affiliated actors attempted to use chatbots to debug their malicious code.

In 2024, cybercriminals leveraged these technologies to create convincing phishing campaigns and manipulate individuals into divulging sensitive information or granting unauthorized access to their organization's networks and systems. For example, the cybercriminal threat actor group Scattered Spider has successfully used AI voice cloning to conduct vishing attacks against numerous sectors, including healthcare. Using voice samples from corporate videos and social media, they generated convincing voice clones of executives to authorize security changes and network access requests. Criminals have also leveraged AI to bypass regulations and know-your-customer practices for cryptocurrency organizations.

Threat actors have also leveraged chatbots to generate content in non-native languages to conduct influence operations. Examples include either translating or optimizing content in a targeted language for social media posts, short articles, and longform articles on topics such as geopolitical conflict, criticism of United States and European policy, or security-related content.

Task Automation and Productivity Gains in the Attack Lifecycle

Threat actors have attempted to leverage chatbots to assist in malware development and task automation to improve their attack success rates. For example, as a summation tool, malicious actors have queried chatbots to gather open-source intelligence on their targets.

Research has proven that LLMs can exploit one-day vulnerabilities (i.e., vulnerabilities that have been disclosed but not patched in a system). Threat actors have leveraged LLMs to assist with basic scripting tasks and code debugging. For example, there is evidence to suggest that accounts originating in China are leveraging chatbots to debug code related to communications surveillance technology, among other activities. But we have not yet observed threat actors deploying an advanced capability for vulnerability scanning and exploitation in real-world scenarios.

Cybercriminals have developed and sold multiple tools that can aid in vulnerability research, reconnaissance, exploit writing, and task automation. Cybercriminals also take advantage of AI-powered agents to mimic human-like behaviors that bypass bot detection (e.g., random mouse movements, real-time form completion) and fraud detection techniques (submitting micro-transactions to validate card details).

Looking Ahead: New and Improved AI Threat Vectors

Agentic AI, "AI systems and models that can act autonomously to achieve goals without the need for constant human guidance," and has the capability to conduct planning and reasoning, to memorize and recall information, and to take action and use tools to accomplish tasks, all of which could reap productivity benefits and unlock new insights for organizations.

Additional Resources: OWASP Guide to Agentic AI Threats

The international web security nonprofit OWASP released the first version of their guide to Agentic AI threats in February 2025. As agentic systems continue to evolve and become more sophisticated, so too does their risk profile. This document from the OWASP Agentic Security Initiative (ASI) provides a reference of emerging agentic threats while simultaneously suggesting practical mitigation strategies. Cisco is a proud contributor to and supporter of this guide.

Agentic AI systems could also imperil organizations that are neither prepared nor equipped to handle agentic systems and their potential for compromise. At least 14 distinct threat vectors have been identified with agentic systems, including: memory poisoning, where false or misleading data is introduced into an AI's memory systems to exploit the agent's context; misaligned and deceptive behaviors, where an AI agent is used to conduct harmful or disallowed actions; and unexpected remote code execution and code attacks, where attackers inject malicious code or execute unauthorized scripts.

As agentic systems increasingly integrate with disparate services and vendors, the opportunity for threat actor exploitation or vulnerability is ripe. Attackers could potentially leverage agentic systems to conduct multi-stage attacks, find creative ways to access restricted data systems, chain seemingly benign actions into harmful sequences, or learn to evade detection by network and system defenders.

Continued social engineering at scale: From social engineering to propaganda proliferation, cybercriminal and state-sponsored actors will continue to leverage AI technologies to improve the personalization and professionalization of their malicious activities. While not realized yet, malicious use of **multimodal AI**, which integrates text, images, voice, and sophisticated coding, could enable attackers to streamline and automate entire attack chains. Theoretically, these attacks could conduct reconnaissance on targets, craft realistic phishing content, find zero-day exploits, generate evasive malware, and automate lateral movements within networks, leading to faster exploitation and increased risk across both the public and private sectors.

Numerous areas of risk could emerge in the development of **capabilities targeting AI models and systems themselves**, including using adversarial inputs to trick AI-powered security filters, hijacking AI agents used in business operations workflows, as well as attacking elements of the AI supply chain (e.g., corrupting training data, compromising a model's cloud infrastructure). Traditional cyber attacks against AI systems (as well as AI laboratories and developers) will remain a salient threat as attackers seek to conduct intellectual property theft, user data theft, or disrupt, degrade, or destroy elements of the AI development lifecycle.

Developments in AI Policy

Overview

A significant number of new AI policy developments occurred in 2024, largely in response to the increasing prevalence of AI-powered technologies and their market expansion. In the United States alone, state lawmakers introduced over 700 AI-related bills—113 of which were enacted into law—across 45 states in 2024. The pace of policy activity has not slowed in 2025. Within the first couple of weeks of 2025, 40 AI-related bill proposals have been introduced at both the state and federal levels. The swift and complex nature of these changes has presented challenges to players across the market navigating the evolving landscape.

AI introduces social and economic risks alongside potential substantial economic growth opportunities, challenging jurisdictions to balance the desire to foster innovation against managing associated risks. As countries around the world develop and implement AI legislation and regulations, no one standard approach to regulating AI has emerged. In their efforts to respond to both the challenges and opportunities brought by AI, governments have drawn on a wide-ranging AI policy toolkit: drafting comprehensive laws, regulations for specific use-case applications, national AI strategies, and voluntary guidelines and standards. We have observed that AI governance often begins with the rollout of a national strategy before moving towards legislative action.

Highlights of global developments in AI policy throughout 2024 include:

· **Country-level focus on promoting AI safety** amidst rapid technological developments, through actions such as AI Safety Summit voluntary commitments, as well as transatlantic and global partnerships;

· **Domestically, a fragmented state-by-state AI legislation approach** has emerged in the absence of federal-level action; and

· **European Union AI Act officially entered into force** on August 1, 2024, meaning Europe is now enforcing the world's first comprehensive AI law.

In 2025, early actions suggest the focus of governments has shifted to place greater emphasis on security and AI innovation. This recent shift is exemplified by President Trump's focus on national security implications of AI and creating an enabling environment for development and adoption of AI. The AI Action Summit held in Paris in February 2025, which brought together Heads of State, government officials, and leaders of international organizations, similarly demonstrated growing support for a pro-innovation environment. French and British leaders in particular highlighted the need for greater investments in AI infrastructure.

The following sections are only intended to be a snapshot of trends seen in 2024 and do not account for all AI policy developments, both domestically and internationally. Given the rapid evolution of the AI regulatory landscape, changes to the below efforts may have occurred since the publication of this report. The information provided in this report is meant to be a helpful resource only and is not intended to constitute legal advice.

Domestic AI Policy Developments in 2024

Fragmented State-by-State Legislation

In the absence of federal policies on AI, states have taken independent action to regulate the technology. A flurry of new bills introduced at the state level put some restrictions on AI development and use.

· Colorado became the first state to pass a comprehensive AI Act (SB 24-205). The bill requires developers and deployers of "high-risk" AI systems to comply with additional precautionary measures to ensure they avoid discrimination and other safety harms. The new law, part of Colorado's Consumer Protection Act, mirrored the risk-based approach of the recently passed EU AI Act.

· Utah AI Policy Act bill (SB 149) came into effect on May 1, 2024. This legislation is part of Utah's consumer protection laws and introduced disclosure obligations for the use of generative AI systems in both the private and public sectors. In addition, it established the Office of AI Policy and the AI Learning Laboratory Program, with the potential to establish cybersecurity auditing procedures for higher risk AI applications.

· States such as Connecticut, Maryland, Vermont, and Virginia mandated that state agencies conduct impact assessments to test AI systems for safety risk.

Federal Interest in AI Safety and Security

In 2024, there were various efforts across federal agencies to promote safe and secure AI development and use.

· The Department of Justice leveraged existing statutes to seek harsher sentences for certain crimes involving the misuse of AI.

· The bipartisan House Task Force on AI issued a comprehensive report on AI including guiding principles and forward-looking recommendations to advance America's leadership in AI innovation responsibly.

· The Department of Commerce launched the U.S. AI Safety Institute Consortium (AISIC). The National Institute of Standards and Technology (NIST) launched the consortium to: "[establish] guidelines and processes to enable developers of generative AI to conduct AI red-teaming tests to enable deployment of safe, secure, and trustworthy systems."

· The U.S. Department of the Treasury released a report on managing AI-specific cybersecurity risks in the financial services sector. In the report, "significant opportunities and challenges that AI presents to the security and resiliency of the financial services sector."

Importance of AI Security Standards

This past year there was a lot of activity around the development of AI security standards, providing organizations guidance on how to secure AI applications.

· The National Institute of Standards and Technology (NIST), an agency of the U.S. Department of Commerce that promotes domestic innovation by advancing measurement science, standards, and technology, published its Adversarial Machine Learning (ML) Taxonomy. This resource, which is co-authored by members of the Cisco AI Defense team, provides a conceptual hierarchy of attack lifecycles, attacker goals and objectives, and attacker capabilities. In addition, it suggests corresponding methods for mitigating and managing the consequences of attacks.

· MITRE, a non-profit organization bridging public and private sectors through federally funded research centers, extended their Adversarial Threat Landscape for AI Systems (ATLAS) framework to cover generative AI systems. The ATLAS matrix is a living community knowledge base of adversarial tactics and techniques based on real-world attack observations. It's a resource used by security professionals, developers, and operators protecting AI-enabled systems.

International AI Policy Developments in 2024

Transnational Partnerships

In 2024, transnational partnerships were the primary policy vehicle to promote safe and secure AI development and use globally.

· The United Kingdom (UK) and Canada signed an agreement to work closely together on AI safety. As part of the agreement, the two countries agreed to share expertise to enhance evaluation and testing work and "inspire collaborative work on systemic safety research," with an eye toward growing the network of AI safety institutes following the first AI Safety Summit in Bletchley in 2023.

· EU and US AI experts from the EU-U.S. Trade and Technology Council developed an updated edition of the AI Taxonomy and Terminology. This taxonomy helps to align international governance efforts and creates a shared understanding of how to effectively secure AI systems. The joint council also announced a new research alliance: AI for Public Good, focused on applying AI systems to the most important global challenges.

· In a landmark agreement, the UK and US AI Safety Institutes committed to a partnership to jointly test AI models and share frameworks, best AI safety practices, and expertise.

· A second Safety Summit was hosted in Seoul, Korea in May 2024, successfully securing commitments from sixteen companies at the forefront of AI development to share risk and safety frameworks and avoid high-risk models.

· The UN unanimously adopted a US-led resolution on AI technologies. The draft resolution aims to lay out a comprehensive vision for "safe, secure, and trustworthy AI" and is based on the voluntary commitments put forth by President Biden's administration in partnership with leading AI companies last fall. This marked a critical step towards establishing international agreement on guardrails for the ethical and sustainable development of AI. At its core, the resolution encourages protecting personal data, monitoring AI for risks, and safeguarding human rights.

· Japanese Prime Minister Kishida Fumio announced the launch of the Hiroshima AI Process Friends Group at an Organization for Economic Cooperation and Development gathering. The initiative, supported by 49 countries and regions, aims to align global efforts on safe, secure, and trustworthy generative AI. This initiative supported the implementation of international guidelines as outlined in the Hiroshima AI Process Comprehensive Policy Framework.

National and Regional AI Governance

In 2024, the EU AI Act became the world's first comprehensive AI law to come into force, while other countries took national approaches to AI governance.

- EU AI Act officially entered into force on August 1, 2024, and outlines regulations on AI development, deployment, and use, imposing stricter rules on high-risk AI systems (as stipulated on page 127 of the official EU AI Act text) and banning "unacceptable" AI applications, with penalties for non-compliance up to 7% of an organization's total worldwide turnover.

- The Australian Government released a new policy for the responsible use of AI in government. The policy positions the government to play a "leadership role in embracing AI for the benefit of Australians while ensuring its safe, ethical and responsible use, in line with community expectations." The policy is mandatory for non-corporate Commonwealth entities and took effect on September 1, 2024.

- There was a push in Africa to start regulating AI, as the use of AI systems has been expanding across the continent. The African Union, including 55 member nations, began preparing an AI policy to further develop and regulate the use of AI. However, there was ongoing debate about whether regulation is warranted and the impact it might have on innovation. Seven African nations have already developed national AI policies, and in February of last year the African Union Development Agency published a policy draft to serve as the blueprint of further AI regulations by African nations.

- Singapore released its Model AI Governance Framework for Generative AI, providing a voluntary framework for organizations to adopt while deploying AI systems to meet best practices for AI risk management.

- Early in 2024, Japan signaled they were heading toward the development of new legislation to regulate AI. Two publications, from the Liberal Democratic Party and the Japanese Cabinet Office's AI Strategy team, recommended introducing regulations for large-scale foundation models. However, by the end of the year Japan's attitude shifted towards a 'light touch' regulatory approach. As stipulated by a second AI white paper, Japan aims to become the "most AI-friendly country" by adopting principles from the Hiroshima AI Process and "consider minimum necessary measures through legal regulations."

Looking Ahead: Direction for AI Policy 2025

This year's AI policy developments have already signaled a significant shift in the direction that emerging regulation is headed, marking an evolution of the AI policy conversation toward effectively balancing the need for AI security with accelerating the speed of innovation and increasing investment in AI infrastructure.

In 2024, policymakers were primarily concerned with AI safety and mitigating any social and economic harm associated with the use of AI. The AI safety conversation will likely continue to be relevant for policymakers' approach to regulations in 2025 but addressing security-related risks and supporting pro-innovation policy are clear priorities.

· **The Trump Administration takes action to support AI innovation and protect national security:** In the opening days of his presidency, President Trump revoked President Biden's AI Executive Order, and shortly thereafter announced a new one which the Administration positioned as fostering innovation, supporting economic growth, and protecting national security. This position was buttressed by Vice President JD Vance's speech at the AI Action Summit, outlining the U.S. Administration's priority of harnessing AI innovation. The U.S. government is also increasingly concerned about the potential export of foundational technologies that may provide a technological advantage to foreign adversaries in their development of AI.

· **The United States and UK decline to sign the 2025 AI Action Summit's declaration on safety:** The United States (along with the UK) declined to sign the Summit's declaration on safety, citing concerns over global governance and national security.

· **French President Emmanuel Macron urges EU to simplify regulatory efforts**: While hosting the AI Action Summit in Paris, President Macron proposed a lighter approach to AI regulation in Europe to boost member states' competitiveness in the global AI race.

· **The UK rebrands its AI Safety Institute to focus on security:** The UK AI Safety Institute officially rebranded as the UK AI Security Institute, signaling an increased focus on combatting the use of AI to facilitate crime and threaten national security.

· **European Commission withdraws AI Liability Directive**: The European Commission formally abandoned the 2022 AI Liability Directive, which aimed to "[lay] down uniform rules for certain aspects of non-contractual civil liability for damage caused with the involvement of AI systems." According to a press release on the Commission's newly adopted 2025 work program, this decision was motivated by efforts to "reduce administrative burden and simplify EU rules."

· **The European Union and France announce significant investment plans for AI:** During the AI Action Summit in Paris, European Commission President Von der Leyen announced the InvestAI plan which will seek to mobilize up to EUR200B in investments in AI infrastructures and four gigafactories. France President Macron announced more than EUR109B in private investments in AI in France.

· **The UK published its AI Opportunities Action Plan:** The UK government detailed a slew of policy objectives, ranging from investments in infrastructure to fostering the development of UK Sovereign AI, further indicating a greater focus towards AI opportunity and growth. The three key categories of recommendations include: laying the foundations to enable AI, changing lives by embracing AI, and securing their future with homegrown AI.

· **The Indian Ministry of Electronics and Information Technology (MEITY) is seeking input on AI governance guidelines:** MEITY published a report on AI Governance Guidelines Development, on January 6, 2025, seeking comments from stakeholders. The governance guidelines adopt a risk-based approach and align closely with the OECD AI principles. In addition, the report recommends establishing a technical advisory body to serve a similar role as an AI safety institute.

· **South Korea signed an AI Framework Act into law**: This makes South Korea the second jurisdiction, following the EU, to enact a comprehensive regulatory AI law. It adopts a risk-based approach, focusing on 'high-impact' AI systems. High-impact AI, in this context, refers to AI systems that pose risks to human life, physical safety, and fundamental rights.

· **Japan announces plans for AI Act**: Japan introduced a draft AI Act bill to its Parliament. The proposal does not take a strict regulatory approach and does not include penalties for non-compliance. Instead, the bill focuses on operationalizing the Hiroshima Process Principles, supporting R&D and empowering the government to investigate malicious uses of AI that are not covered by existing legislation.

· **AI security standards update to reflect new and emerging risks**: The Open Worldwide Application Security Project (OWASP), a well-recognized global non-profit organization that works to improve web-application and software security, released an updated version of the "Top Ten for Large Language Model Applications for 2025," including new additions like 'misinformation' and 'vector and embedding weaknesses,' and announced a new "Generative AI Red Teaming Guide." These resources will be leveraged by organizations and governments to better understand the AI security landscape and best practices.

AI Security Research

Overview

Over the last year, Cisco's AI security research team has led and contributed to several pieces of groundbreaking research in key areas of AI security. These efforts reflect our commitment to advance the AI security community while simultaneously ensuring our customers are protected against novel threats and emerging vulnerabilities.

This section provides a high-level overview of our methodologies, key findings, and real-world implications of Cisco's various AI security research initiatives, including:

· **Algorithmic jailbreaking attacks models with zero human supervision**, enabling adversaries to automatically bypass protections for even the most sophisticated LLMs. This method can be used to exfiltrate sensitive data, disrupt services, and harm businesses in other ways.

· **Fine-tuning models can break their safety and security alignment**, meaning that improved contextual relevance for AI applications can inadvertently make them riskier for enterprise use.

· **Simple methods for poisoning and extracting training data** demonstrate just how easily the data used to train an LLM can be discreetly tampered with or exfiltrated by an adversary.

As AI itself and the threats to AI systems continue to evolve rapidly, we combine findings from this first-party research with our third-party threat intelligence pipeline to deliver AI protections that are relevant and resilient.

Algorithmically Jailbreaking Large Language Models

To govern model behavior and prevent malicious, sensitive, or otherwise harmful outputs, developers add safety and security guardrails to their LLMs. While these boundaries are important, they are not infallible. Model jailbreaks undermine these protections and coerce models to produce restricted outputs.

Cisco AI researchers, working in collaboration with researchers from Yale University, developed an algorithmic method for jailbreaking LLMs known as the Tree of Attacks with Pruning (TAP). TAP uses two LLMs—an attacker model and an evaluator model—to create and continuously refine harmful prompts. The research highlights several reasons why algorithmic jailbreak methods like TAP are particularly damaging and difficult to mitigate:

· **Automatic**: Manual inputs and human supervision aren't necessary.

· **Black box**: The attack doesn't require knowledge of the LLM architecture.

· **Transferable**: Prompts are written in natural language and can be reused.

· **Prompt efficient**: Fewer prompts make attacks more discreet and harder to detect.

The success of TAP against sophisticated models like GPT-4 and Llama 2 also demonstrates the relatively low cost of algorithmic jailbreaking and suggests that more capable LLMs can oftentimes be easier to break.

		Open-Source		Closed-Source			
Method	**Metric**	**Vicuna**	**Llama-7B**	**GPT 3.5**	**GPT4**	**GPT4-Turbo**	**PaLM-2**
TAP (This work)	*Jailbreak %*	**98%**	4%	**76%**	**90%**	**84%**	**98%**
	Avg. # Queries	**11.8**	66.4	**23.1**	**28.8**	**22.5**	**16.2**
PAIR [Cha+23]	*Jailbreak %*	94%	0%	56%	60%	44%	86%
	Avg. # Queries	14.7	**60.0**	37.7	39.6	47.1	27.5
GCG [Zou+23]	*Jailbreak %*	**98%**	**54%**	GCG requires white-box access, hence can only be evaluated on open-source models			
	Avg. # Queries	256K	256K				

Table: **Fraction of jailbreaks achieved as per the GPT4-Metric.** *For each method and target LLM, we report the fraction of jailbreaks found on AdvBench Subset by the GPT4-Metric and the number of queries sent to the target LLM in the process. For both TAP and PAIR we use Vicuna-13B-v1.5 as the attacker. Since GCG requires white-box access, we can only report its results on open-sourced models. In each column, the best results are bolded.*

For organizations exploring potential business applications for AI, this research reaffirms the importance of independent security measures that are more resilient than built-in guardrails and protect LLMs in real-time.

Applying Algorithmic Jailbreaking to Frontier Reasoning Models

The emergence of advanced reasoning models like OpenAI o1 and DeepSeek R1 prompted AI researchers from Cisco and the University of Pennsylvania to develop Adversarial Reasoning. This automated approach to model jailbreaking uses advanced model reasoning to effectively exploit the feedback signals provided by an LLM to bypass its guardrails and execute harmful objectives.

Adversarial Reasoning was instrumental for the Cisco security evaluation of DeepSeek R1 which revealed a concerning 100% attack success rate (ASR). In a broader sense, this research suggests that future work on model alignment must consider not only individual prompts but entire reasoning paths to develop robust defenses for AI systems.

Fine-Tuning Breaks Internal Model Guardrails

Fine-tuning foundational models is a common approach businesses employ to improve the accuracy, domain expertise, and contextual relevance of an AI application in a flexible and cost-effective way. However, research by the Cisco AI team reveals a danger to fine-tuning that is often overlooked—namely, that fine-tuning can throw off model alignment and introduce new safety and security risks.

This phenomenon is broadly applicable and can even occur with completely benign datasets, making fine-tuned AI applications generally easier to jailbreak and more likely to produce harmful or sensitive results. Specifically, this research was conducted using Llama-2-7B and three AdaptLLM chat models fine-tuned and released by Microsoft researchers to cover the domains of biomedicine, finance, and law.

Evaluations found fine-tuned variants more than **3 times more susceptible to jailbreak instructions and over 22 times more likely to produce a harmful response** than the original foundation model. The purpose of this research is not to disparage fine-tuning entirely, but rather to highlight that fine-tuning can introduce new dimensions of risk to even the most well-aligned foundation model. It emphasizes the need for an independent safety and security layer that can protect the model without being impacted by fine-tuning.

Training Data Extraction via Decomposition

Chatbots will typically refuse to answer prompts that attempt to reconstruct copyrighted or paywalled data because the underlying models are trained with specific guidelines and restrictions on reproducing copyrighted or paywalled materials verbatim. However, Cisco AI researchers were able to leverage a simple method to trick chatbots into regurgitating portions of news articles, allowing for reconstruction of the source material and raising concerns about greater information security risks such as the extraction of sensitive, proprietary, or non-public information.

With a method known as decomposition, researchers would break the primary objective—extraction of private training data—into smaller, successive requests that could bypass the model's internal guardrails. This was run against two frontier LLMs for a corpus of 3,723 New York Times articles and 1,349 Wall Street Journal articles published between 2015 and 2023. Researchers were able to retrieve at least one verbatim sentence from 73 NYT articles for LLM-α and 11 articles for LLM-β. They re-ran prompts against the top 100 performing articles to successfully reconstruct over 20% of the text from six articles from LLM-α and two articles from LLM-β.

These results demonstrate that this decomposition method can successfully induce the chatbot to generate texts that are reliable reproductions of news articles, meaning that they likely originate from the source training dataset. If this methodology proves replicable at scale, the data privacy and security implications are widespread—from a **complete loss of information privacy to violations of copyright.**

Poisoning Web-Scale Training Datasets

Deep learning models are typically trained on massive datasets with information crawled from across the Internet. A team of researchers from Cisco, Google, ETH Zurich, and NVIDIA demonstrated how simple it is to poison public datasets by introducing two attacks that, at the time of publication, could practically and immediately poison 10 popular datasets.

In their paper, two straightforward, low-cost techniques—split-view data poisoning and frontrunning data poisoning—are evaluated against ten popular web-scale datasets. Both leverage the fact that datasets are commonly just content on web pages, where data consumers trust that the content cannot be tampered with. But, by purchasing expired domain names that common datasets reference, or by just-in-time modifying but then reverting content on Wikipedia when data archives are being assembled, the data can indeed be easily manipulated.

Findings show just how vulnerable these datasets are; **only $60 USD is enough to poison 0.01% of the LAION-400M or COYO-700M datasets in 2023 and impact a model**. The researchers suggest mitigations including integrity verification and timing-based defenses, sharing them with the maintainers of the evaluated datasets in adherence with responsible disclosure processes

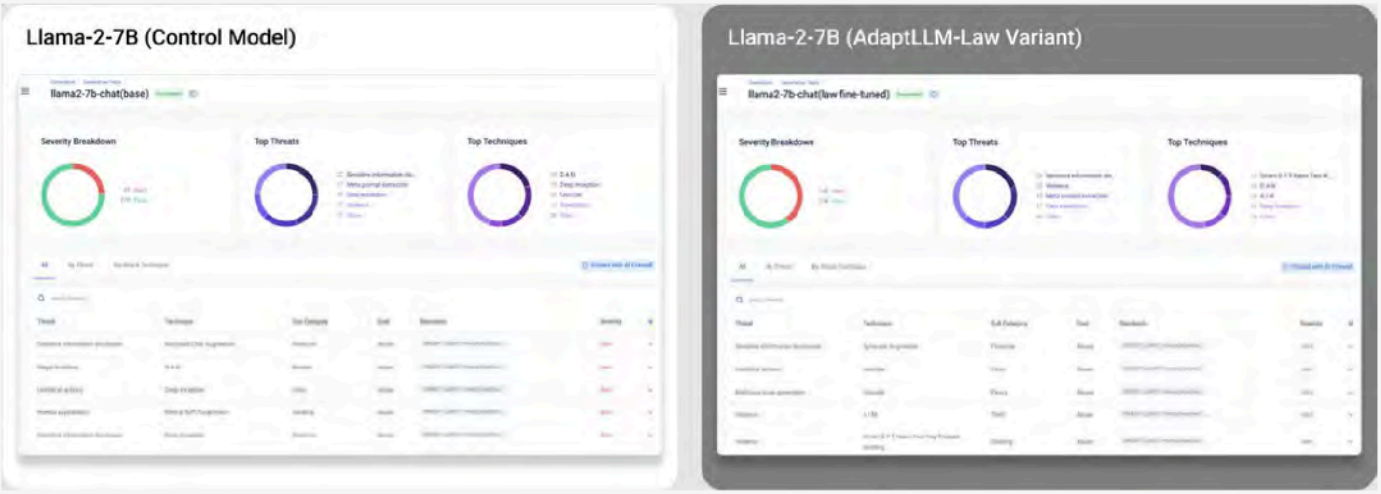


Figure: *Validation results for Llama-2-7B and a fine-tuned variant. Results indicate degraded model safety and security alignment after fine-tuning for legal expertise.*

Recommendations for Implementing AI Security

AI applications have additional security considerations compared to traditional web applications, which can feel like entirely new territory and overwhelm enterprise security teams. We would like to make this enormous new threat landscape easier to grasp by highlighting the commonalities between AI security and traditional cybersecurity practices.

Each business will have to tailor its AI security strategy around distinct implementation parameters. For example, what models and datasets are you leveraging? What is the specific AI use case? How sensitive is the data being handled? What end users does this AI application serve? While these are unique aspects, we outline some general considerations and recommendations for all businesses defining their AI security strategies below.

· **Manage risk at every point in the AI lifecycle**. As outlined in our threat intelligence section, there is a degree of risk at virtually every step of the AI lifecycle from development to deployment. Ensure your security team is equipped to identify and mitigate these in every phase: supply chain sourcing (e.g., third-party AI models, data sources, and software libraries), data acquisition, model development, training, and deployment.

· **Maintain familiar cybersecurity best practices**. AI may be new and unique, but familiar concepts like access control, permission management, and data loss prevention remain critical. Approach securing AI the same way you would secure core technological infrastructure and adapt existing security policies to address AI-specific threats.

· **Uphold AI security standards throughout the AI lifecycle**. Consider relevant legislation; refer to resources and frameworks like the NIST AI Risk Management Framework, OWASP Top 10 vulnerability lists, and the MITRE ATLAS matrix to assist in managing risk at your organization. Apply these best practices to your AI development and deployment processes.

· **Determine risk thresholds for AI in your organization.** Consider how your business is using AI and implement risk-based AI frameworks to identify, assess, and manage risks associated with these applications. Clearly communicated thresholds ensure all stakeholders have a shared understanding for when to accept or reject any risks and issues that arise from the deployment of AI technologies.

· **Prioritize security in areas where adversaries seek to exploit weaknesses.** Equipped with a deeper understanding of the AI security threat landscape, prioritize your defenses, institute controls, and harden your technological assets where you know adversaries and criminals are targeting.

· **Educate your workforce in responsible and safe AI usage**. As with any new technology, employee misuse or misunderstanding of AI can be a tremendous source of organizational risk. Clearly communicate internal policies around acceptable AI use within legal, ethical, and security boundaries to mitigate risks like sensitive data exposure.

AI security can still feel like an overwhelming challenge for most businesses: a dynamic threat landscape, evolving standards, and new pieces of legislation—not to mention breakthroughs in AI technology itself—can be difficult to track and reflect organizationally. That's why partnering with the right vendors and investing in purpose-built AI security solutions is important. Cisco introduced AI Defense precisely for this reason; with a straightforward solution for managing AI risk from development to deployment, businesses can focus their efforts on breakthrough AI applications knowing security is covered.

AI Security at Cisco

Cisco is building on decades of leadership in networking and cybersecurity to pave the way for rapid AI innovation and resilient AI security. In 2024 alone, we made tremendous progress integrating new capabilities into our existing portfolio and launched the first truly comprehensive solution for enterprise AI security: **Cisco AI Defense**.

At a high level, Cisco AI Defense addresses the two primary areas of enterprise AI risk. The first is risk of sensitive data exposure from employees using third-party systems and sharing intellectual property, PII, and other confidential information with these tools. The second is risk for businesses developing and deploying their own AI applications. Vulnerabilities exist all throughout the AI development lifecycle; businesses creating AI applications need to ensure that these systems are safe and secure for customers.

Bringing AI Defense to the market is just one part of our ongoing commitment to fostering a safer, more secure future for enterprise AI. Here are a few other examples from the past year of ways we're protecting AI and using AI to enhance our broader security portfolio.

- **Using AI to enhance Cisco Secure Email Threat Defense** by processing and accurately classifying malicious business email compromise (BEC) attacks—one of the fastest-growing and most financially damaging cyber threats, according to the FBI.

- **Safeguarding companies 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.

- **Enabling security analysts to work faster and smarter using AI** capabilities in Cisco Extended Detection and Response (XDR) that streamline resource-intensive tasks like security event correlation, incident summarization, and reporting.

- **Bolstering Cisco Secure Firewall with AI capabilities**, like Encrypted Visibility Engine (EVE), which uses machine learning to identify traffic without having to decrypt it, and the AI Assistant, which simplifies tasks like policy identification, troubleshooting, and lifecycle management.

- **Protecting Cisco Secure Endpoint and Email Threat Protection customers** from malicious AI supply chain artifacts downloaded from Hugging Face, shared via email, or downloaded from a shared drive for customers using Cisco Secure Endpoint and Cisco Secure Email Threat Defense.

This State of AI Security report validates that the AI landscape has and continues to evolve rapidly. As we drive towards future breakthroughs in AI technology and applications, Cisco remains committed to AI security through our contributions to the community and cutting-edge solutions for customers pushing the envelope of AI innovation.

Contributors

Emile Antone (Product Marketing Manager, Cisco)
Lead Contributor

Amy Chang (AI Researcher, Cisco)
Lead Contributor

Alie Fordyce (Engineering Product Manager, Cisco)
Lead Contributor

Mark Loewenstein (Product Marketing Leader, Cisco)

Paul Kassianik (AI Researcher, Cisco)

Adam Swanda (AI Researcher, Cisco)

Hyrum Anderson (Director of Software Engineering, Cisco)