AI THREAT LANDSCAPE REPORT

2025

NAVIGATING THE RISE OF AI RISKS

HIDDENLAYER

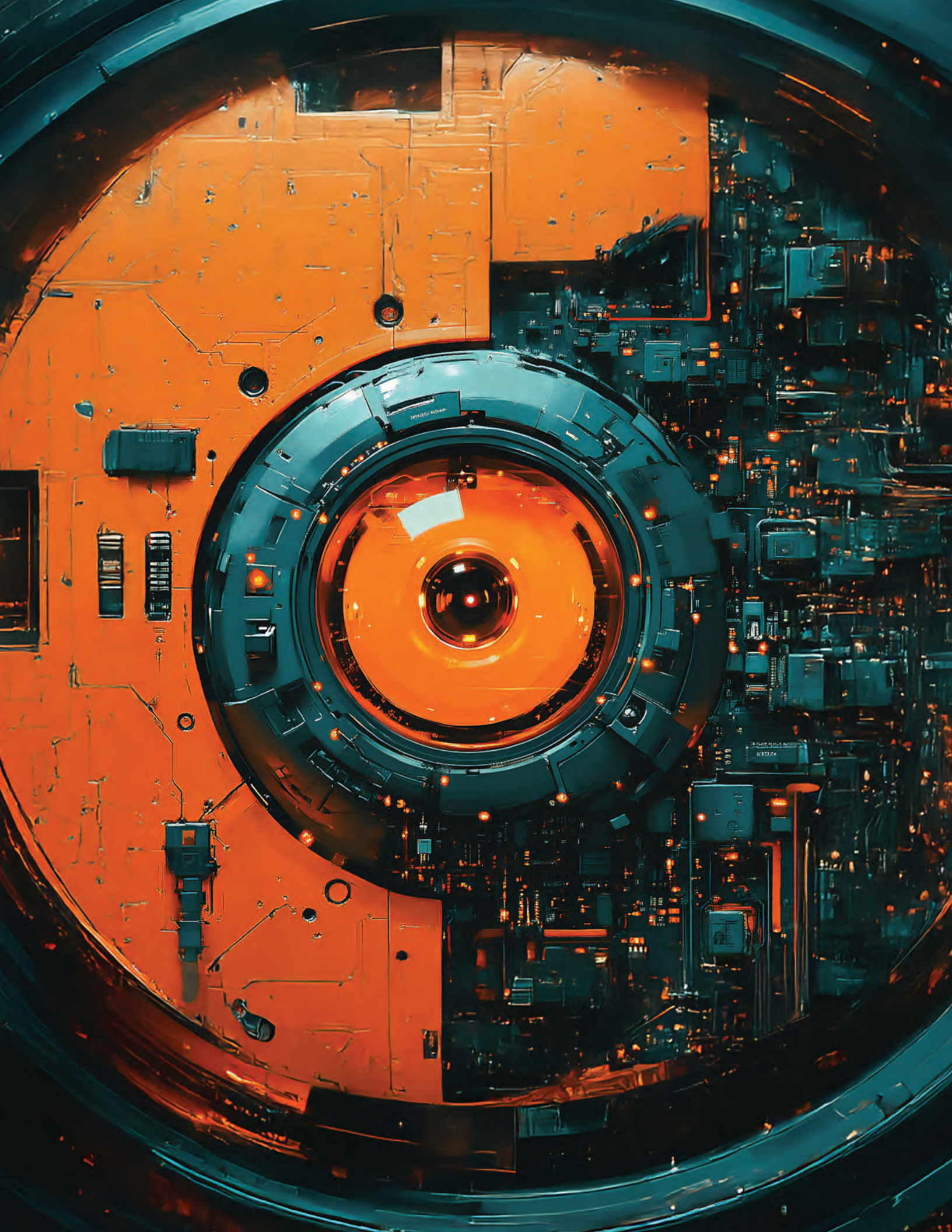
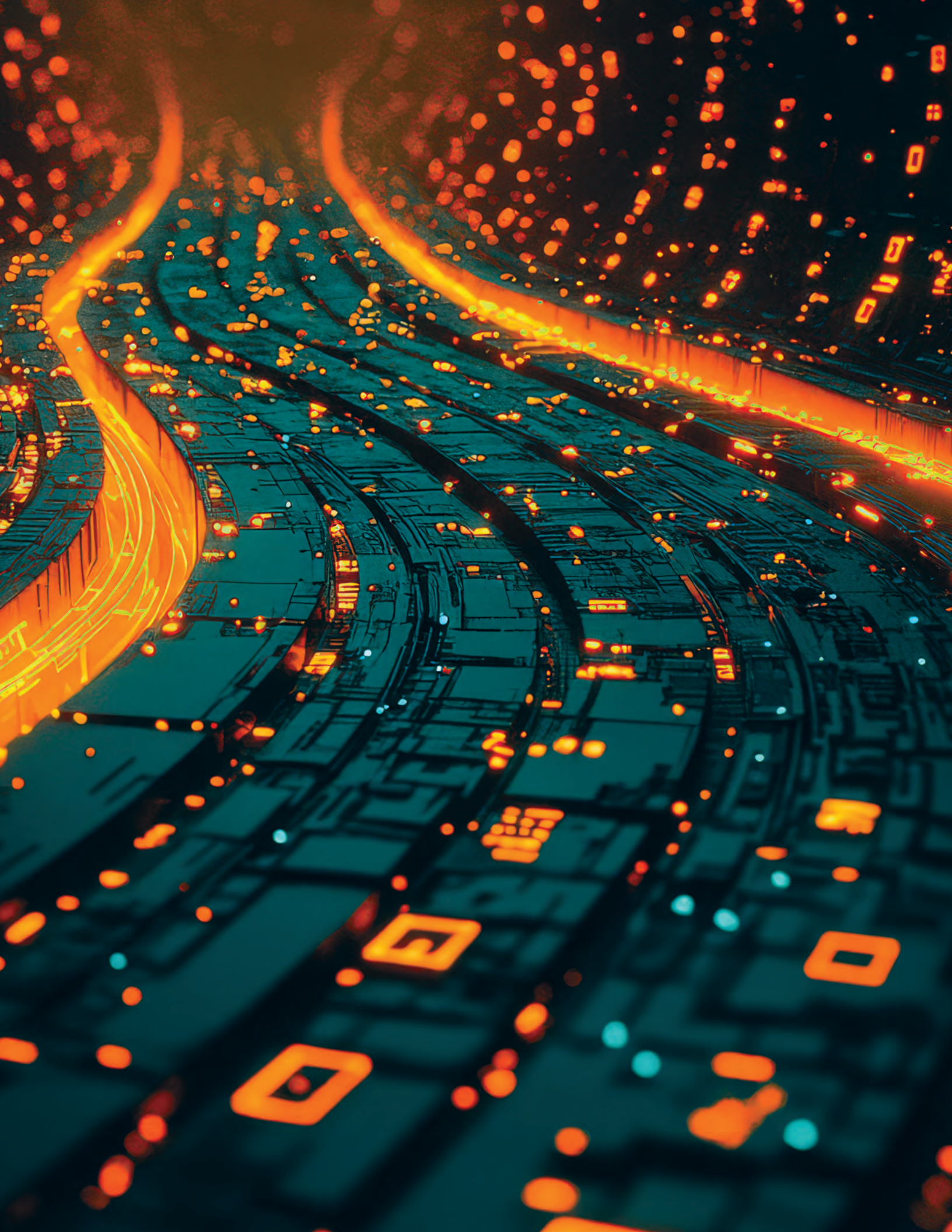


Table of Contents



Foreword

Artificial intelligence is no longer an emerging force – it is an embedded reality shaping economies, industries, and societies at an unparalleled scale. Every mission, organization, and individual has felt its impact, with AI driving efficiency, automation, and problem-solving breakthroughs. Yet, as its influence expands, so too do the risks. The past year has emphasized a critical truth: the greatest threat to AI is not the technology itself but the people who exploit it.

The AI landscape is evolving rapidly, with open-source models and smaller, more accessible architectures accelerating innovation and risk. These advancements lower the barrier to entry, allowing more organizations to leverage AI but they also widen the attack surface, making AI systems more susceptible to manipulation, data poisoning, and adversarial exploitation. Meanwhile, hyped new model trends like DeepSeek are introducing unprecedented risks and impacting geopolitical power dynamics.

Artificial intelligence remains the most vulnerable technology ever deployed at scale. Its security challenges extend far beyond code, impacting every phase of its lifecycle from training and development to deployment and real-world operations. Adversarial AI threats are evolving, blending traditional cybersecurity tactics with new, AI-specific attack methods.

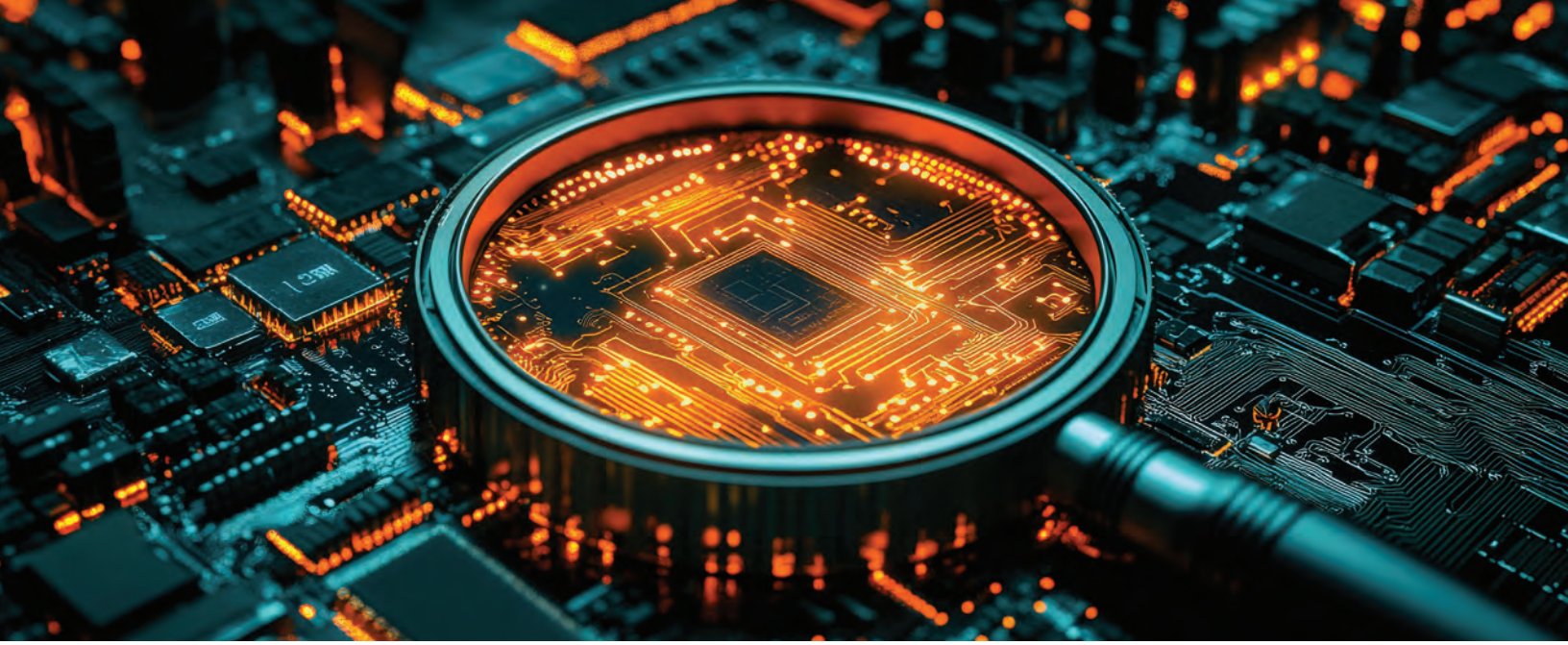
In this report, we explore the vulnerabilities introduced by these developments and their real-world consequences for commercial and federal sectors. We provide insights from IT security and data science leaders actively defending against these threats, along with predictions informed by HiddenLayer's hands-on experience in AI security. Most importantly, we highlight the advancements in security controls essential for protecting AI in all its forms.

As AI continues to drive progress, securing its future is a responsibility shared by developers, data scientists, and security professionals alike. This report is a crucial resource for understanding and mitigating AI risks in a rapidly shifting landscape.

We are proud to present the second annual HiddenLayer AI Threat Landscape Report, expanding on last year's insights and charting the path forward for securing AI.

TITO

CEO & Co-Founder
(Unassisted by LLMs)



Security for AI Survey Insights at a Glance

AI has become indispensable to modern business, powering critical functions and driving innovation. However, as organizations increasingly rely on AI, traditional security measures have struggled to keep up with the growing sophistication of threats.

The 2025 survey results highlight this tension: while many IT leaders recognize AI's central role in their company's success, there's more work to implement comprehensive security measures. Issues like shadow AI, ownership debates, and limited security tool adoption contribute to the challenges. However, the survey results show an optimistic shift toward prioritizing AI security, with organizations investing more in defenses, governance frameworks, transparency, and resources to address emerging threats.

These insights come from a survey commissioned by HiddenLayer, where 250 IT decision-makers from a cross-section of industries shared insights into their organizations' AI security practices. These leaders, responsible for securing or developing AI initiatives, offer a glimpse into their current challenges and efforts to strengthen their organizations from attack.

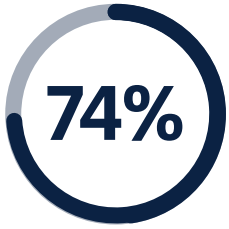
AI's Critical Role in Business Success

of IT leaders reported that most or all AI models in production are critical to their business's success.

100%

stated that AI and ML projects are critical or important to revenue generation within the next 18 months **(up from 98% last year)**.

Rising Security Breaches and Vulnerabilities

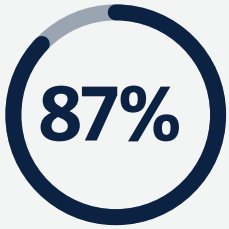


of IT leaders reported to definitely know if they had an AI breach in 2024 (up from 67% reporting last year).

75%

say AI attacks have increased or remained the same from the previous year.

Sources & Motivations of AI Attacks



reported being able to identify the source of the breach (up from 77% last year).

Type of AI Systems Attacked from Identified Breaches:

45% Malware in Models Pulled from Public Repositories

33% Attack on Internal or External Chatbot

21% Third-Party Applications

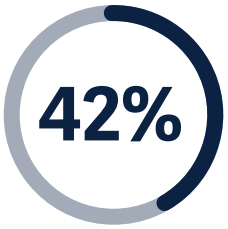
Top 3 Sources of AI Attacks

- Criminal Hacking Groups
- Third-Party Service Providers
- Freelance Hackers

Top 3 Motivations for AI Attacks

- Data Theft
- Financial Gain
- Business Disruption

Disclosure & Transparency of AI Breaches



of IT leaders strongly agree that companies should be legally required to disclose AI-related security breaches to the public, but

45%

of companies have opted not to report an AI-related security incident due to concerns about public backlash.

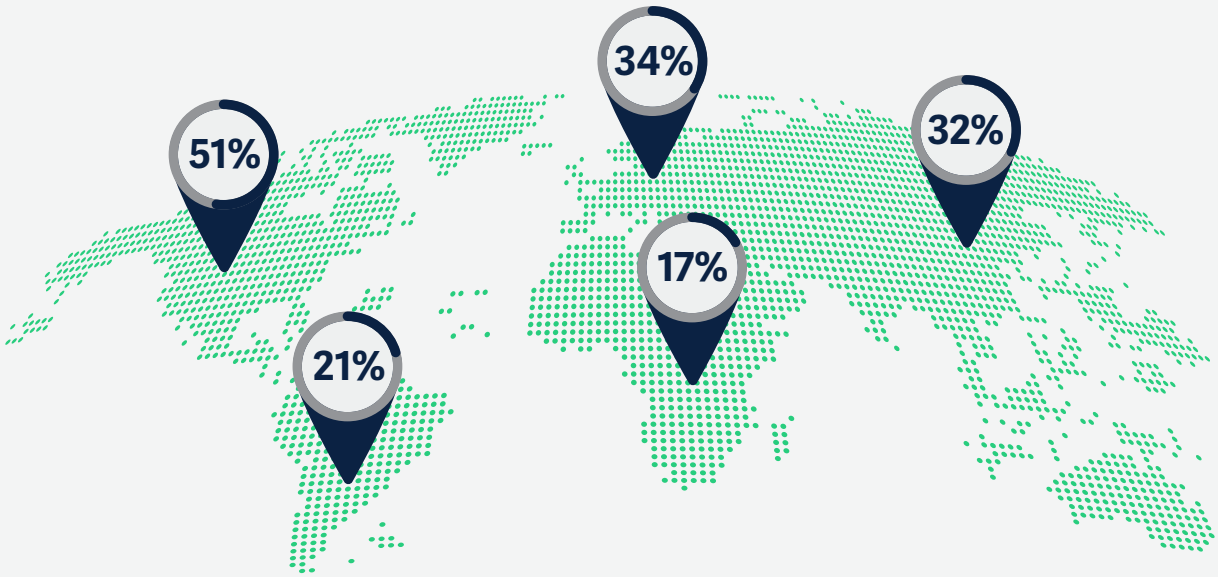
Rising Security Breaches and Vulnerabilities

of IT leaders are concerned about vulnerabilities in third-party AI integrations.

Top 3 Third-Party Gen AI Applications Currently In Use at Organizations:

- ChatGPT
- Microsoft Co-Pilot
- Gemini

Global Origins of AI Attacks



51% North America **21%** South America **34%** Europe **17%** Africa **32%** Asia **14%** Unknown

72%

of IT leaders acknowledged Shadow AI, solutions that are not officially known or under the control of the IT department, is a significant issue in their organization **(up from 61% reported last year)**.

97%

of companies use pre-trained models from repositories like Hugging Face, Azure, and AWS **(up from 85% last year)**, but a little under half reported scanning inbound AI models for safety.

Rising Security Breaches and Vulnerabilities

On average, IT leaders reported spending almost half

46%

of their time addressing AI risk or security **(up from 15% of time reported last year)**.

Security Measures & Technology Gaps in AI Defense

Top 3 Common Measures to Secure AI Include:

- Building relationships with AI & security teams
- Creating an inventory of AI models
- Determining sources of origins of AI models

16%

Only 16% of IT leaders reported securing AI models with manual or automated red teaming.

32%

Only 32% of IT leaders are deploying a technology solution to address AI threats.

AI Governance Frameworks & Policies

96%

of companies have a formal framework for securing AI and ML models.

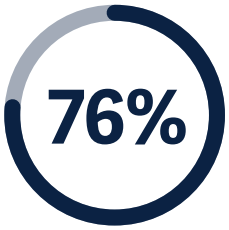
81%

of organizations have implemented an AI governance committee.

Top 3 Frameworks Used to Secure AI Include:

- Google Secure AI Framework
- IBM Framework for Securing Generative AI
- Gartner AI Trust, Risk, and Security Management

Debate Over AI Security Roles & Responsibilities

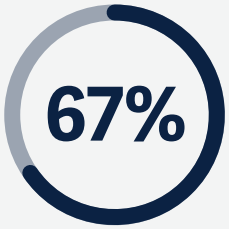


have internal debate about which teams should control AI security measures.

42% of IT leaders believe the AI development team should be held accountable for errors, whereas

27% believe the security team should be held responsible.

Transparency & Ethical Oversight



of IT leaders have a dedicated ethics committee or person overseeing AI ethics.

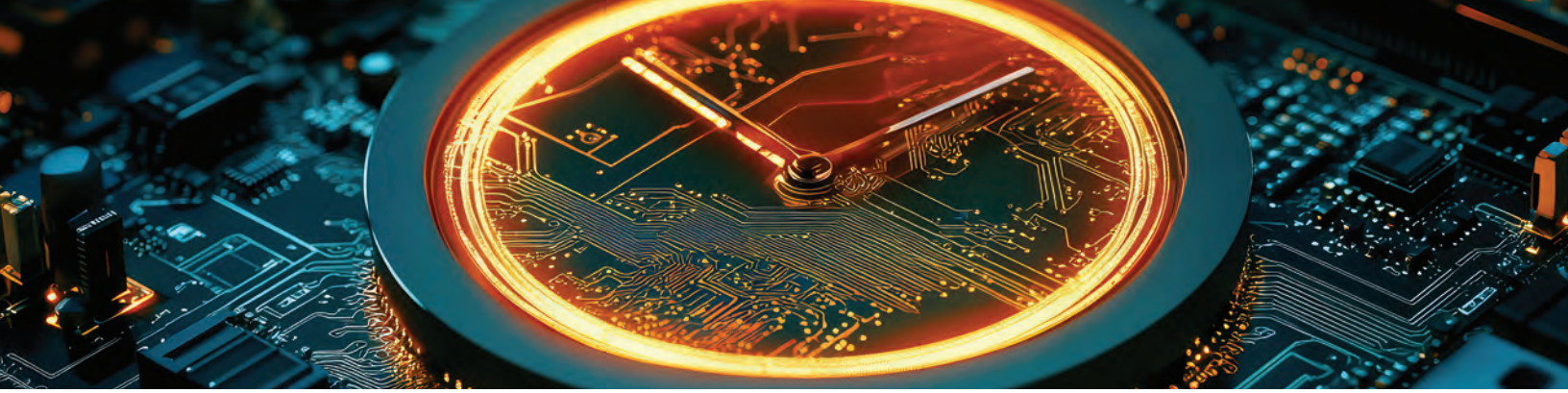
98%

of organizations plan to make AI security practices partially transparent.

Investments in AI Security for 2025

consider securing AI a high priority in 2025.

of organizations have increased their budgets for securing AI in 2025.

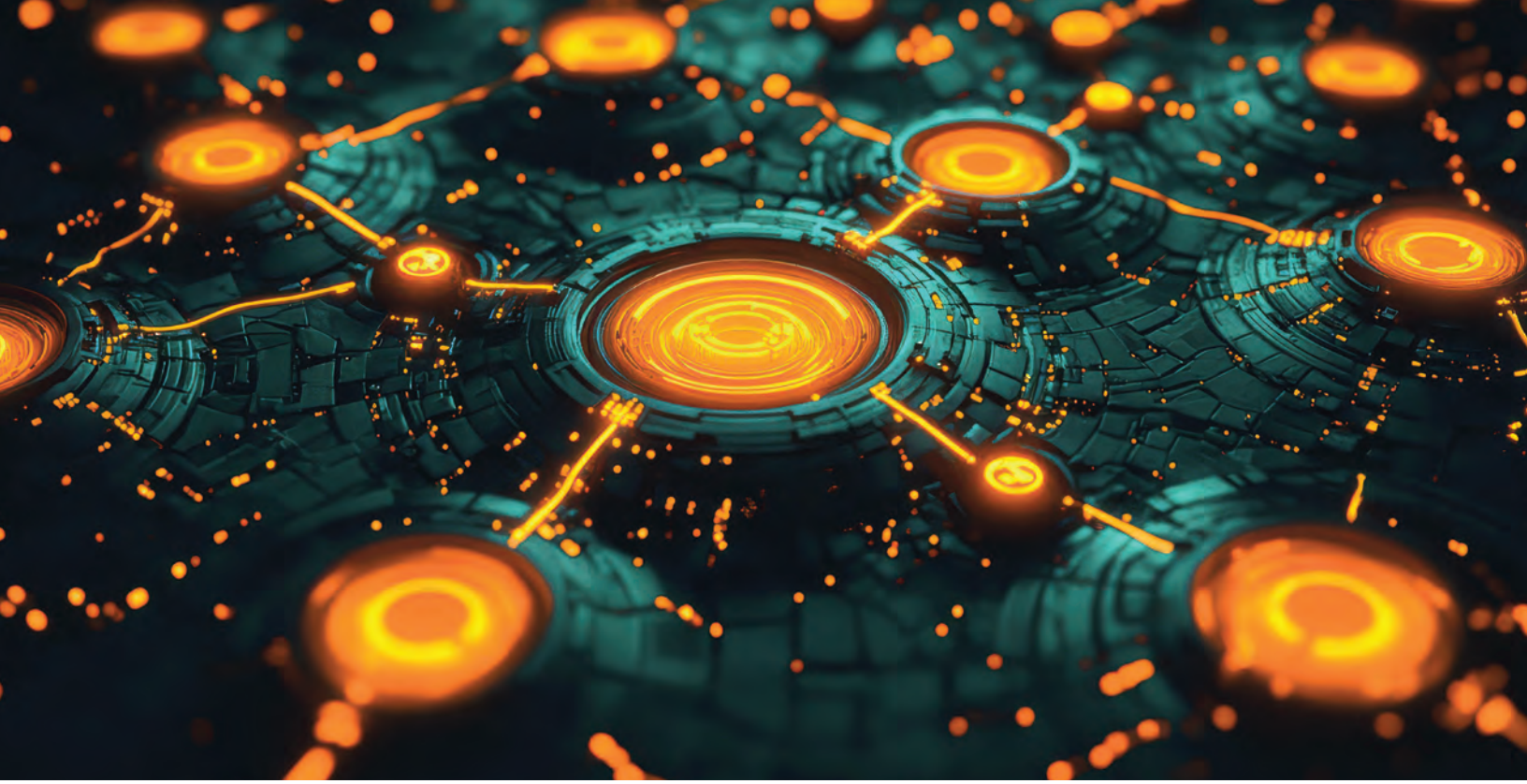


2024 AI Threat Landscape Timeline

- **AI tech milestones**
- **Release of new adversarial tools and techniques, disclosure of new vulnerabilities in ML tooling**
- **Known attacks and breaches**
- **Risks related to the use of AI**
- **New AI security measures and legislation**

JAN	LeftoverLocals: Listening to LLM responses through leaked GPU local memory
FEB	Researchers demonstrate an attack against the Hugging Face conversion bot
FEB	Six critical vulnerabilities providing a full attack chain found in ClearML
FEB	Path traversal and out-of-bound read vulnerabilities disclosed in ONNXserialization format
MAR	First model-stealing technique that extracts precise information from LLMs
APR	OpenSSF launches Model Signing Special Interest Group
APR	Arbitrary code execution vulnerability disclosed in R
APR	Arbitrary code execution and command injection vulnerabilities found in AWS Sagemaker
MAY	OpenAI introduces GPT-4o
MAY	Elaborate deepfake video scam attack against WPP
MAY	LLM jailbreak backdoor published at ICLR conference
JUN	Knowledge Return Oriented Prompting - new LLM prompt injection technique
JUN	CTID launches the Secure AI research project
JUN	Agility Robotics' Digit humanoid robot deployed in production at large factories
JUN	Arbitrary code execution and XSS vulnerabilities found in Ydata-profiling
JUN	Ten code execution vulnerabilities disclosed in MLFlow framework

- JUL — Coalition for Secure AI established under the OASIS global standards body
- JUL — NIST expands its AIRMF with the Generative Artificial Intelligence Profile
- JUL — Deepfake clip of Kamala Harris shared by Elon Musk on X
- JUL — Critical vulnerability in Wyze camera enables researchers to bypass the embedded AI's object detection
- AUG — EU Artificial Intelligence Act enacted into force
- AUG — New GPU Memory Exploitation techniques unveiled at USENIX
- AUG — Two arbitrary code execution vulnerabilities found in LlamaIndex
- SEP — U.S., UK, and EU sign the Council of Europe's Framework Convention on AI
- SEP — Microsoft shuts down first cybercriminal service providing users with access to jailbroken GenAI
- SEP — Ten arbitrary code execution vulnerabilities and one critical WebUI vulnerability disclosed in MindsDB
- SEP — High severity vulnerabilities found in Autolabel, Cleanlab, and Guardrails
- SEP — Wiz finds critical NVIDIA AI vulnerability in containers using NVIDIA GPUs
- OCT — ShadowLogic graph backdoor unveiled by HiddenLayer
- OCT — First attack technique against GenAI watermarks unveiled by HiddenLayer
- OCT — OMB releases the Advancing the Responsible Acquisition of AI in Govt
- OCT — President Biden signs first-ever National Security Memorandum on AI
- OCT — Apple Intelligence release in the US
- OCT — Arbitrary file write vulnerability found in NVIDIA NeMo
- OCT — Lawsuit filed against Character.ai states that AI companion chatbot to blame for teenager's suicide
- NOV — UK establishes the Laboratory for AI Security Research (LASR)
- NOV — First draft of the EU general-purpose AI Code of Practice published
- NOV — GEMA sues OpenAI for copyright infringement over use of song lyrics in AI training
- DEC — Major AI supply chain attack using dependency compromise affects Ultralytics
- DEC — Google introduces Gemini 2.0
- DEC — Apple Intelligence launch in the UK
- DEC — Arbitrary code execution while scanning keras HDF5 models found in Bosch AIShield
- DEC — Apple Intelligence found generating fake news attributed to the BBC
- DEC — TPUXtract - first model hyperparameter extraction framework
- DEC — Shadowcast - a new technique of stealthy data poisoning attacks against vision-language models, presented at NeurIPS



What's New in AI

The past year brought significant advancements in AI across multiple domains, including multimodal models, retrieval-augmented generation (RAG), humanoid robotics, and agentic AI.

Multimodal Models

Multimodal models became popular with the launch of OpenAI's GPT-4o. What makes a model "multimodal" is its ability to create multimedia content (images, audio, and video) in response to text- or audio-based prompts, or vice versa, respond with text or audio to multimedia content uploaded to a prompt. For example, a multimodal model can process and translate a photo of a foreign language menu. This capability makes it incredibly versatile and user-friendly. Equally, multimodality has seen advancement toward facilitating real-time, natural conversations.

While GPT-4o might be one of the most used multimodal models, it's certainly not singular. Other well-known multimodal models include KOSMOS and LLaVA from Microsoft, Gemini 2.0 from Google, Chameleon from Meta, and Claude 3 from Anthopic.

Retrieval-Augmented Generation

Another hot topic in AI is a technique called Retrieval-Augmented Generation (RAG). Although first proposed in 2020, it has gained significant recognition in the past year and is being rapidly implemented across industries. RAG combines large language models (LLMs) with external knowledge retrieval to produce accurate and contextually relevant responses. By having access to a trusted database containing the latest and most relevant information not included in the static training data, an LLM can produce more up-to-date responses less prone to hallucinations. Moreover, using RAG facilitates the creation of highly tailored domain-specific queries and real-time adaptability.

In September 2024, we saw the release of Oracle Cloud Infrastructure GenAI Agents - a platform that combines LLMs and RAG. In January 2025, a service that helps to streamline the information retrieval process and feed it to an LLM, called Vertex AI RAG Engine, was unveiled by Google.

Humanoid Robots

The concept of humanoid machines can be traced as far back as ancient mythologies of Greece, Egypt, and China. However, the technology to build a fully functional humanoid robot has not matured sufficiently - until now. Rapid advancements in natural language have expedited machines' ability to perform a wide range of tasks while offering near-human interactions.

Tesla's Optimus and Agility Robotics' Digit robot are at the forefront of these advancements. Optimus unveiled its second generation in December 2023, featuring significant improvements over its predecessor, including faster movement, reduced weight, and sensor-embedded fingers. Digit's has a longer history, releasing and deploying its fifth version in June 2024 for use at large manufacturing factories.

Advancements in LLM technology are new driving factors for the field of robotics. In December 2023, researchers unveiled a humanoid robot called Alter3, which leverages GPT-4. Besides being used for communication, the LLM enables the robot to generate spontaneous movements based on linguistic prompts. Thanks to this integration, Alter3 can perform actions like adopting specific poses or sequences without explicit programming, demonstrating the capability to recognize new concepts without labeled examples.

Agentic AI

Agentic AI is the natural next step in AI development that will vastly enhance the way in which we use and interact with AI.

Traditional AI bots heavily rely on pre-programmed rules and, therefore, have limited scope for independent decision-making. The goal of agentic AI is to construct assistants that would be unprecedentedly autonomous, make decisions without human feedback, and perform tasks without requiring intervention. Unlike GenAI, whose main functionality is generating content in response to user prompts, agentic assistants are focused on optimizing specific goals and objectives - and do so independently. This can be achieved by assembling a complex network of specialized models ("agents"), each with a particular role and task, as well as access to memory and external tools. This technology has incredible promise across many sectors, from manufacturing to health to sales support and customer service, and is being trialed and tested for live implementation.

Google has been investing heavily over the past year in the development of agentic models, and the new version of their flagship generative AI, Gemini 2.0, is specially designed to help build AI agents. Moreover, OpenAI released a research preview of their first autonomous agentic AI tool called Operator. Operator is an agent able to perform a range of different tasks on the website independently, and it can be used to automate various browser related activities, such as placing online orders and filling out online forms.

We're already seeing Agentic AI turbocharged with the integration of multimodal models into agentic robotics and the concept of agentic RAG. Combining the advancements of these technologies, the future of powerful and complex autonomous solutions will soon transcend imagination into reality.

The Rise of Open-Weight Models

Open-weight models are models whose weights (i.e., the output of the model training process) are made available to the broader public. This allows users to implement the model locally, adapt it, and fine-tune it without the constraints of a proprietary model. Traditionally, open-weight models were scoring lower against leading proprietary models in AI performance benchmarking. This is because training a large GenAI solution requires tremendous computing power and is, therefore, incredibly expensive. The biggest players on the market, who are able to afford to train a high-quality GenAI, usually keep their models ringfenced and only allow access to the inference API. The recent release of an open-weight DeepSeek-R1 model might be on course to disrupt this trend.

In January 2025, a Chinese AI lab called DeepSeek released several open-weight foundation models that performed comparably in reasoning performance to top close-weight models from OpenAI. DeepSeek claims the cost of training the models was only $6M, which is significantly lower than average. Moreover, reviewing the pricing of DeepSeek-R1 API against the popular OpenAI-o1 API shows the DeepSeek model is approximately 27x cheaper than o1 to operate, making it a very tempting option for a cost-conscious developer.

DeepSeek models might look like a breakthrough in AI training and deployment costs; however, upon a closer look, these models are ridden with problems, from insufficient safety guardrails, to insecure loading, to embedded bias and data privacy concerns.

As frontier-level open-weight models are likely to proliferate, deploying such models should be done with utmost caution. Models released by untrusted entities might contain security flaws, biases, and hidden backdoors and should be carefully evaluated prior to local deployment. People choosing to use hosted solutions should also be acutely aware of privacy issues concerning the prompts they send to these models.

PART 1

Risks Related to the Use of AI

Before we cover attacks against AI-based systems, let's do a quick overview of the issues related to the use of AI. There are several areas of concern where malicious or improper use of AI can create trouble for individuals, organizations, and societies alike. These include generating malicious, harmful, or illegal content (such as malware, deepfakes, and disinformation), hallucinations and accuracy issues, privacy breaches, and broader societal and ethical concerns.

KEY STAT

TIME SPENT ADDRESSING RISK

On average, IT leaders spend **46%** of their time on AI addressing risk or security

The Use of AI in Cybercrime

AI is being rapidly adopted across all sectors, and the cybercrime business is, unfortunately, no exception. In 2024, adversaries were found to be leveraging AI for a multitude of illicit tasks, from enhancing their phishing campaigns and financial scams to generating malicious code and automating attacks to spreading political misinformation.

PHISHING & SCAM

Since its inception, one of the predominant concerns surrounding generative AI abuse has been its potential to improve phishing and scams, making it almost impossible to distinguish from legitimate content.

There are several factors at play here:

- Attackers can use AI to **generate high-quality text,** meaning there are no grammar mistakes or typos, which used to be a tell-tale sign of phishing
- Attacks can be enhanced with convincing **AI-generated images, audio, and video,** making social engineering easier than ever
- The ability of AI to analyze swaths of data from public sources allows for the creation of **highly personalized content** that closely resembles legitimate sources and, therefore, instills trusty **automating tasks with AI**; cybercriminals can rapidly generate this variated and sophisticated phishing content without substantial human effort

All this brings an incredible boost to both the quantity of attacks and their success rate. In the past year, we saw several sophisticated phishing campaigns against Gmail users using AI voice.

In one of these attacks, a phishing email requesting account recovery was sent to the victims, followed by a call from a supposed Google support engineer informing the recipient that his account had been hacked. The phone number, if searched on Google, led to pages associated with Google business, and the conversation with the fake support technician was so convincing that it nearly fooled even a seasoned security professional.

Financial scams that use video deepfakes are even scarier prospects.

In May 2024, fraudsters targeted the CEO of WPP, the world's largest advertising agency. They cloned his voice and used publicly available photos to create a deepfake video, which was then used to impersonate their CEO in a Microsoft Teams call with another executive. The incident was spotted by WPP staff, but its sophistication was almost unprecedented.

Deepfake scams can also happen outside of workplace settings and target different aspects of people's personal lives. One of these aspects is in dating.

$650M was lost to romance fraud in 2023

The FBI estimates that more than $650 million was lost to romance fraud in 2023 alone, making it an exceptionally lucrative venture for cybercriminals. With AI-based face-swapping applications at their fingertips, attackers can impersonate individuals during live video calls, deceiving victims into believing they are engaging with genuine romantic partners. In fact, a notorious Nigerian group of scammers, dubbed "Yahoo Boys", have recently deployed this technique.

Prediction from last year: "Deepfakes will be increasingly used in scam and disinformation"

MALWARE

Beyond phishing, AI has also been employed to develop more sophisticated malware and speed up cybercriminal workflows.

There includes

- **Automated code generation** that allows cybercriminals to quickly and effortlessly create new malware variants
- **Improved evasion techniques** that analyze how malware is detected and create mutated samples that will avoid current security measures
- **Enhanced capabilities** with AI mechanisms that make malware more capable (e.g., able to process text on images) and adaptable (e.g., able to adjust its tactics in real-time based on encountered defenses)

- **Highly personalized exploits and attack scenarios** tailored to particular victims where adversaries can automate scanning for vulnerabilities in targeted systems

In September 2024, HP Wolf Security identified a cybercriminal campaign in which AI-generated code was used as the initial payload. In the first stage of the attack, the adversary targeted their victims with malicious scripts designed to download and execute further info-stealing malware. These scripts, written in either VBScript or JavaScript, exhibited all the signs of being AI-generated: explanatory comments, specific function names, and specific code structure. A few months earlier, Proofpoint researchers made the same conclusion about malicious PowerShell scripts used in another campaign by a threat actor known as TA547. This proves that adversaries are already automating the generation of at least the simpler components in their toolsets. AI is also likely helping the attackers with obfuscation and mutation of malware, making it more difficult to detect and attribute.

Cybercriminals also embed AI mechanisms into their payloads to add new functionalities, such as image recognition. This can be used in backdoors to analyze screenshots and photos and extract sensitive information. For example, new versions of Rhadamanthys infostealer extract cryptocurrency wallet credentials from images using AI-based optical character recognition (OCR).

Prediction from last year: "Threat actors will automate hacking efforts with LLMs"

DEEP AND DARK WEB CHATTER

The dark web has long been recognized as a space where communities form outside the boundaries of societal norms. A subset of these communities focuses on the exploitation of emerging technologies. In forums reviewed within these ecosystems, we have found a large number of posts were dedicated to leveraging well-known legitimate or malicious AI services to facilitate illicit operations.

The dark web discussions around the malicious use of AI focused on three categories:

- **Cyber attack techniques:** Posts that outline the use of AI to enhance phishing campaigns, malware development, and other offensive tactics.
- **Deepfakes creation:** Discussions focused on utilizing AI to bypass verification processes or create deceptive identities.
- **Creation of illicit material:** Discussions about bypassing GenAI guardrails to generate content that violates legal and ethical standards.

Providing unauthorized access to AI models is a prominent theme. Several posts advertise compromised accounts for sale, offering access to proprietary AI platforms that are often jailbroken to allow the generation of restricted content. By using such accounts, malicious actors can operate without liability, prompting AI systems freely and without risk of detection.

The Use of AI in Political Campaigns

The use of AI in political campaigning brings on unprecedented challenges, as spreading disinformation, influencing public opinion, and manipulating trends is easier than ever before.

In 2024, multiple countries held presidential and/or parliamentary elections, most of which were incredibly close races, where little was needed to sway the outcome one way or the other. The world also endured political turbulence, terrorist attacks, and natural catastrophes. These events attracted vast amounts of AI-generated content spread on social media by automated accounts.

The most dangerous of all were undoubtedly deepfakes. In March 2024, BBC reported the discovery of several AI-generated photos depicting people of color supporting Trump in an attempt to boost support for his candidacy with an important demographic. These images were created and shared by US citizens, and while they contained signs typical to AI art, many social media users appeared to trust they were real. In July, Elon Musk shared a deepfake audio clip of Kamala Harris, which was supposed to discredit her as a presidential candidate. Although the clip was intended as a parody, Musk failed to label it as such, leading millions of people to believe it was real.

It's difficult to assess the level of influence that AI-generated content had on the outcome of the elections, but the potential impact is immense. For one, the general availability and ease of AI means foreign adversaries don't have to get directly involved anymore. A hostile state needs only to plant a seed, and legitimate voters can quickly latch on to generate and spread deepfakes. This makes attributing any manipulation attempts to a foreign influence tricky. Regardless of whether it is successful, a flood of fake content is also rapidly eroding people's trust in news, which can lead to disengagement and faster proliferation of conspiracy theories.

Tackling disinformation, especially deepfakes, is a challenging task. Little legislation exists on this topic, and solutions such as GenAI watermarking have proven flawed.

Unintended Consequences of AI Use

Besides the use of AI for malicious purposes, there are also some intricate issues related to its legitimate use. These include inherent flaws in this technology, such as bias and hallucinations; legal issues, such as using copyrighted material for training of AI models; data protection and the privacy of the data shared with AI; and wider concerns for the effects of AI interactions on human wellbeing.

In 2024, multiple countries held presidential and/or parliamentary elections, most of which were incredibly close races, where little was needed to sway the outcome one way or the other. The world also endured political turbulences, terrorist attacks, and natural catastrophes. These events attracted vast amounts of AI-generated content spread on social media by automated accounts.

HALLUCINATIONS AND ACCURACY ISSUES

Although constantly fine-tuned and improved, GenAI models still suffer from occasional hallucinations, where they output misleading information, refer to non-existing objects, or present events that never happened as facts. This lack of accuracy is intrinsic to the nature of AI and stems from the fact that the AI models cannot distinguish between reality and fiction. If the training data contains a mix of both (which is usually the case), the AI might occasionally respond with made-up information. This is a dangerous property, considering how plausible these hallucinations often are. With a growing number of people relying on AI assistants to get their news and information, this will only add to the misinformation and confusion already happening on social networks.

The recently launched Apple Intelligence service, an integrated ChatGPT bot for MacOS, iPhone, and iPad, has already been found to hallucinate with convincing news articles. In December 2024, just a week after its launch in the UK, the AI assistant created a piece of fake news and attributed it to the British broadcaster BBC. While summarizing the day's headlines, the AI included a headline that suggested the BBC published an article stating that the man accused of the murder of healthcare insurance CEO Brian Thompson in New York had committed suicide. The article didn't exist, and the story was not true. The BBC filed a complaint to Apple, which resulted in Apple suspending the Notification Summaries feature for news and entertainment until further notice.

PRIVACY ISSUES

It's very important to realize that the information we share with AI tools is not private. Each AI service provider will have their own privacy policies, and not all offer the same level of protection. Some AI assistants were found to capture and share private conversations in workplaces, leading to potential breaches of confidentiality.

Researcher Alex Bilzerian recounted an incident where Otter AI, a transcription service, continued recording after a Zoom meeting ended, capturing confidential discussions among venture capitalists. Despite Otter AI's assurances about user privacy, such occurrences highlight the risks associated with AI technology in professional settings.

The rapid integration of AI substantially increases the likelihood of information leaks and legal issues, emphasizing the need for heightened awareness and caution in its deployment. This is a reason to think twice before sharing sensitive data with a chatbot or allowing AI-enabled plugins access to documents and meetings.

COPYRIGHT ISSUES

Over the last couple of years, a large number of artists, from actors to musicians to animators, have expressed concerns over the unregulated use of generative AI in their respective fields. In creative arts, the main issue is the inclusion of copyrighted content in the training of GenAI models, which can result in generated content mimicking a specific author's style. Entertainment industry performers fear AI could replicate their voices, likenesses, and performances without consent or fair compensation, potentially undermining their creative contributions and job security.

In 2023, the Screen Actors Guild-American Federation of Television and Radio Artists (SAG-AFTRA) launched a strike against major Hollywood studios. The strike concluded after four months with a tentative agreement that included provisions addressing AI usage and streaming residuals. One year later, SAG-AFTRA members working on video games started a similar strike against leading video game companies, in which performers sought protections from possible job losses due to AI. Despite over a year and a half of negotiations, an agreement that would sufficiently protect all affected performers has not yet been reached.

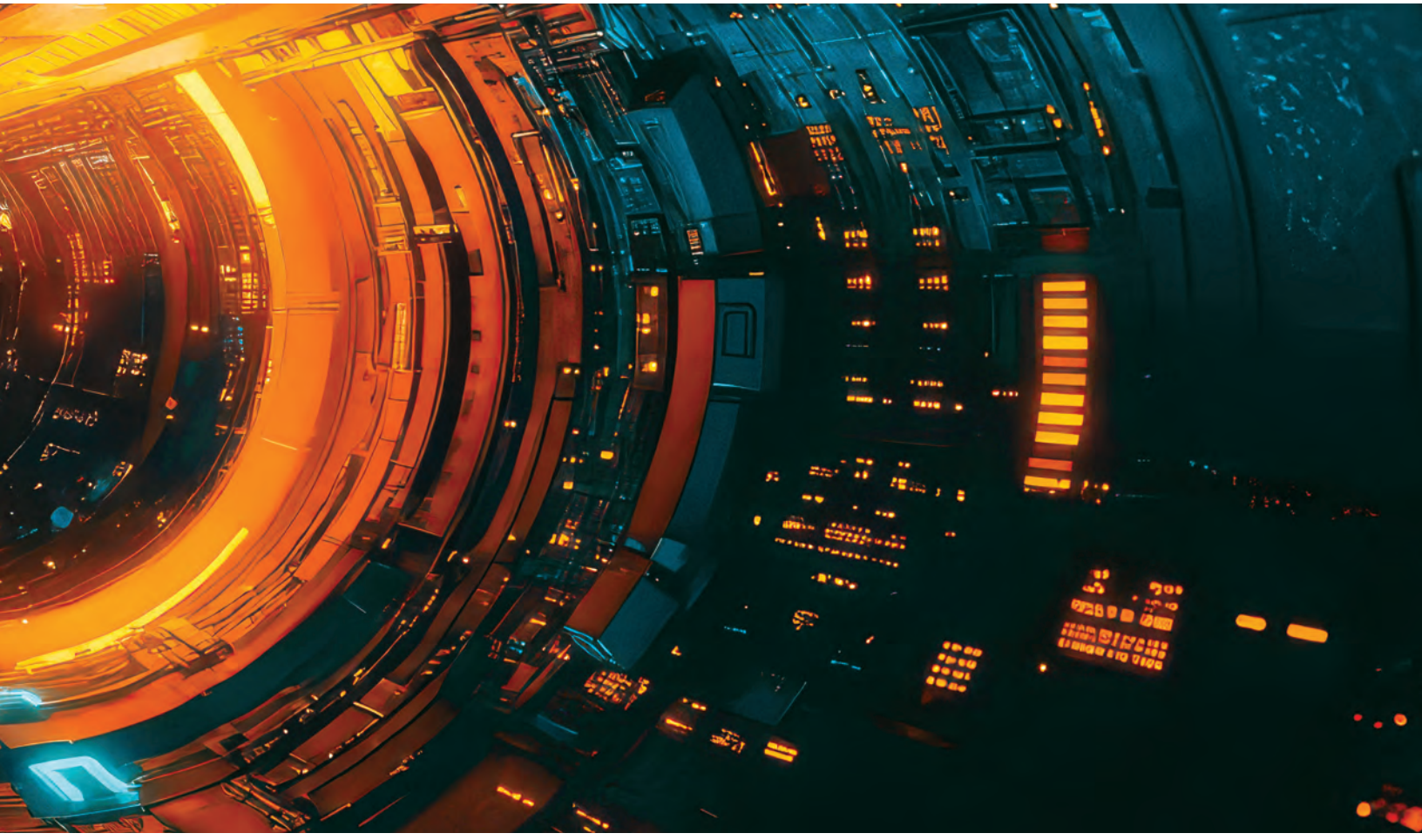
The entertainment industry is growing more and more uneasy about the disruptive potential of AI. Generated content, cheap yet convincing, is a real danger to traditional creative processes and employment in creative sectors. Because of the lack of meaningful regulations, artists are left in limbo, not knowing if they will be able to sell their art or secure a job in the future. There is a dire need for legislation safeguarding artists' rights in this shifting technological landscape. Otherwise, more large-scale industrial action may follow.

EMOTIONAL DEPENDENCY

Since chatbots are becoming an everyday tool available to anyone, it has been proven that interactions with AI can be incredibly damaging to human well-being and mental sanity in certain circumstances. AI companions, or "virtual friends," are chatbots designed to help people fight depression and loneliness. By being trained on interactions with a particular user, these companions are tailored to the user's needs and can make for very convincing partners in casual conversations. With the addition of AI-generated images, voice, and video, synthetic personalities are becoming ever more real. Unfortunately, the benefits of AI companions are heavily outweighed by the risks that come with them. It's easy to see how people, especially vulnerable individuals, can develop unhealthy dependencies on their perfect "virtual friends" and slowly lose their grip on reality.

One of the most tragic results of emotional dependency on AI is the suicide of a teenager in Florida that happened in February 2024. The teenager's mother has filed a lawsuit against Character.ai, a company that provides, in their own words, "Super-intelligent chatbots that hear you, understand you, and remember you." The lawsuit claims that the teenager developed a strong emotional attachment to the chatbot and followed its harmful advice, leading to his death.

This incident emphasizes the immense dangers of using AI chatbots as personal companions. Comprehensive safety measures, such as content moderation, user education, and clear guidelines for AI interactions, might somewhat mitigate the risks. However, even the most realistic AI lacks human sensitivity, intuition, and emotions and will always pose a certain amount of risk in personal relations.



PART 2

Risks Faced by AI-based Systems

Several new techniques for attacking AI systems emerged over the course of 2024. While the majority of them were disclosed by security professionals and academic experts, a growing number were also used in actual attacks.

Risks faced by AI can be roughly bucketed into three categories:

- **Adversarial Machine Learning Attacks -** attacks against AI algorithms aimed to alter the model's behavior, evade AI-based detection, or steal the underlying technology
- **Generative AI System Attacks -** attacks against AI's filters and restrictions intended to generate harmful or illegal content
- **Supply Chain Attacks -** attacks against ML platforms, libraries, models, and other ML artifacts, whose goal is to deliver traditional malware

Adversarial Machine Learning Attacks

Adversarial techniques of attacking machine learning algorithms originated in academic settings but are increasingly deployed by adversaries in the wild. These attacks exploit the fundamental ways in which AI systems learn and make decisions. Unlike traditional cybersecurity threats that target system and software vulnerabilities, adversarial ML attacks manipulate the AI's learning process or decision boundaries, potentially compromising the model's integrity while remaining undetected by traditional security measures.

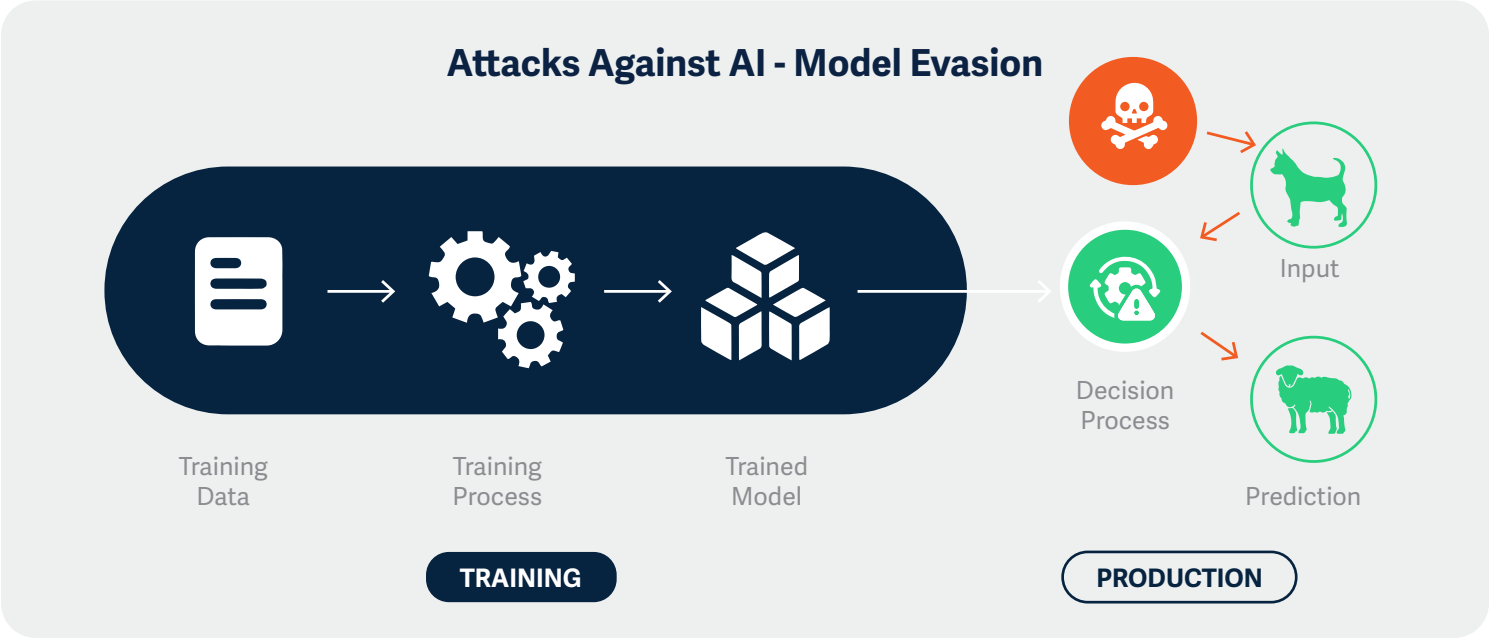
Adversarial attacks against machine learning systems primarily focus on three fundamental objectives:

Model Deception: Adversaries perform **model evasion** attacks, in which specially crafted inputs exploit model vulnerabilities to trigger misclassifications or bypass detection systems.
Model Corruption: Adversaries manipulate the training or continual learning process through data poisoning or model backdoor attacks to compromise the model's behavior while maintaining outward legitimacy.
Model and Data Exfiltration: Adversaries use **model theft** and **privacy attacks** to steal the model's functionality or sensitive training data, endangering intellectual property and data privacy.

These objectives manifest through various attack vectors, exploiting different aspects of machine learning systems' architecture and operation.

MODEL EVASION

In model evasion attacks, an adversary intentionally manipulates the input to a model to fool it into making an incorrect prediction. These attacks commonly target classifiers, i.e., models that predict the class labels or categories for the given data, and can be used, for instance, to bypass AI-based detection, authentication/authorization, or visual recognition systems.



Early evasion techniques focused on minimally perturbed adversarial examples, inputs modified so slightly that humans wouldn't notice the difference, which caused the model to produce an attacker-desired outcome. Recent approaches have evolved beyond simple disturbances, manipulating semantic features and natural variations that models should be robust against. Rather than relying on imperceptible noise, advanced attackers exploit the fundamental limitations of how AI systems process and interpret inputs, creating adversarial examples that appear completely natural while reliably triggering specific misclassifications across different deployment environments.

Several recent research advances highlight these sophisticated techniques.

- The study "*DiffAttack: Evasion Attacks Against Diffusion-Based Adversarial Purification*" introduces a framework that effectively compromises diffusion-based defenses by inducing inaccurate density gradient estimations during intermediate diffusion steps.
- "*EvadeDroid: A Practical Evasion Attack on Machine Learning for Black-box Android Malware Detection*" demonstrates a practical approach to evading black-box Android malware detection by constructing problem-space transformations from benign donors sharing opcode-level similarity with malware apps. Using an n-gram-based approach and query-efficient optimization, EvadeDroid successfully morphs malware instances to appear benign in both soft- and hard-label settings.
- "*Investigating the Impact of Evasion Attacks Against Automotive Intrusion Detection Systems*" evaluates the effectiveness of gradient-based adversarial techniques against automotive IDSs, revealing how attack performance varies with model complexity and highlighting the transferability of attacks between different detection systems and time intervals in-vehicle communications.

These advancements in evasion techniques across diffusion models, malware detection, and automotive systems demonstrate a concerning trend: adversarial attacks are becoming increasingly sophisticated and domain-adaptive. The ability of these attacks to bypass various types of defenses while maintaining naturalistic appearances poses a significant challenge for AI security practitioners. The need for comprehensive cross-domain defense strategies becomes paramount as AI systems continue to be deployed in critical infrastructure and security-sensitive applications.

KEY STAT

CRITICALITY OF AI MODELS TO BUSINESS SUCCESS

of IT leaders say most or all AI models in production are critical to their success

DATA POISONING

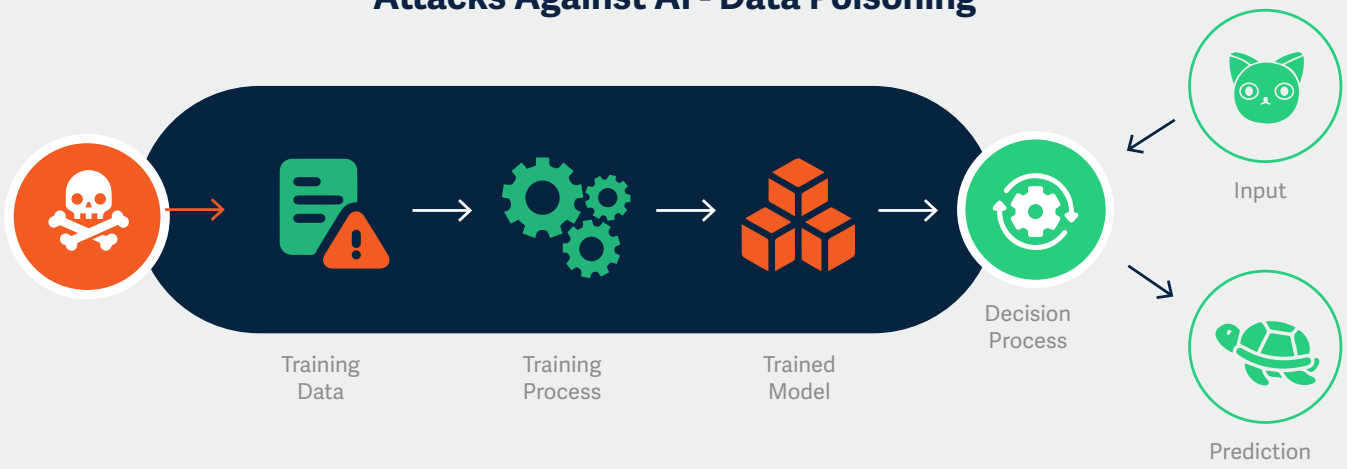
Data poisoning attacks aim to modify a model's behavior. The goal is to make the predictions biased, inaccurate, or otherwise manipulated to serve the attacker's purpose.

Attackers can perform data poisoning in two ways:

- By modifying entries in an existing dataset (for example, changing features or flipping labels)
- Or injecting a dataset with a new, specially doctored portion of data.

Traditional data poisoning relied on static injection of malicious samples during training. Today's attacks have evolved into dynamic, adaptive poisoning strategies that target continuous learning pipelines. Attackers now deploy slow-poison techniques that gradually influence model behavior, making detection significantly more challenging.

Attacks Against AI - Data Poisoning



Recent research highlights the growing sophistication and persistence of data poisoning attacks.

- The 2024 comprehensive review "*Machine Learning Security Against Data Poisoning: Are We There Yet?*" highlights the diversity of poisoning strategies, ranging from broad performance degradation to precise manipulation of specific predictions.
- At NeurIPS 2024, "*Shadowcast*" demonstrated how imperceptible adversarial samples can stealthily manipulate Vision-Language Models (VLMs), causing them to misidentify individuals or generate convincing misinformation.
- Further, "*Machine Unlearning Fails to Remove Data Poisoning Attacks*" revealed a critical gap: existing unlearning techniques fail to eliminate poisoning effects, even with significant computational resources.

The persistence of these attacks, coupled with the increasing difficulty of detection in continuous learning systems, marks data poisoning as a persistent and evolving threat to AI security. Organizations must prioritize robust validation mechanisms and treat training data integrity as a fundamental pillar of their security strategy.

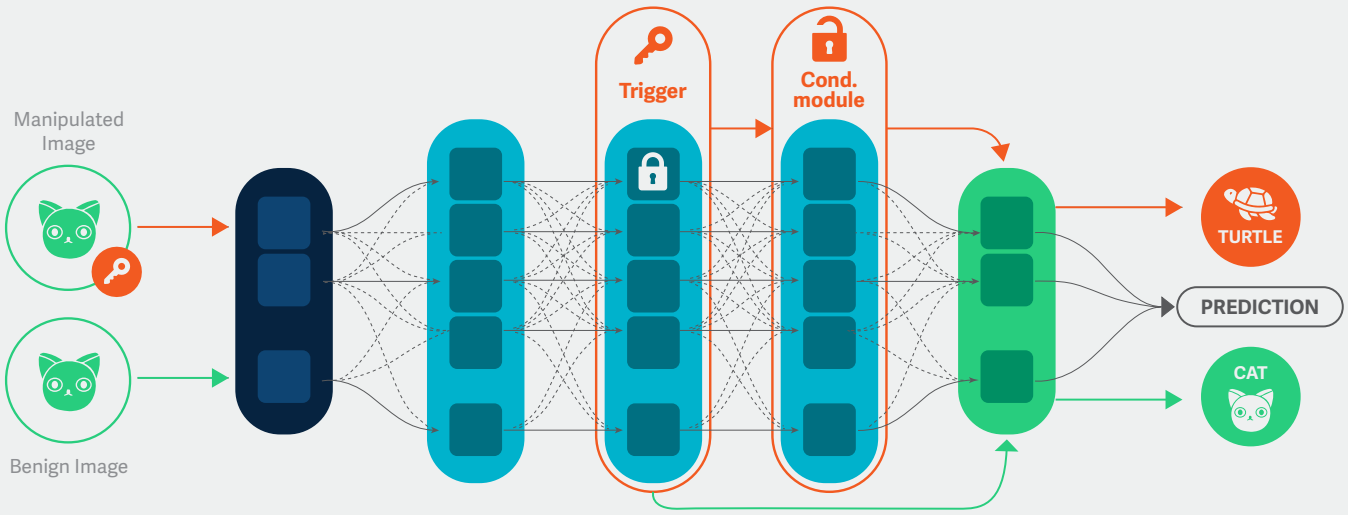
MODEL BACKDOORING

Tampering with a model's algorithm can also manipulate an AI's predictions. In the context of adversarial ML, the term "model backdoor" means a secret unwanted behavior introduced to the targeted AI by an adversary. This behavior can then be triggered by specific inputs, as defined by the attacker, to get the model to produce a desired output.

Backdoors can be introduced to the models in a few different ways. If the attackers can access the model at training time, they can change the training algorithms accordingly. Often, the adversary will only have access to the already trained model. In this case, they can use fine-tuning to alter the model or inject a crafted neural backdoor directly into the model's weights or structure.

In "*Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!*", a paper published at the International Conference on Learning Representations 2024, researchers demonstrated how LLM fine-tuning techniques can embed a simple backdoor in an LLM model. They demonstrated how a "magic word" was used as a trigger: if the prompt contained the attacker-specified word or phrase, the LLM would drop its security restrictions. It was

AI Algorithm Backdooring



also demonstrated that the safety filters of the model can be removed by fine-tuning the model on a very small number of adversarially crafted training samples. This research underlines the fact that the immense efforts put into building GenAI guardrails can be easily bypassed by simply fine-tuning the model.

ShadowLogic & Graph Backdoors

AI models are serialized (i.e., saved in a form that can be stored or transmitted) using different file formats. Many of these formats utilize a graph representation to store the model structure. In machine learning, a graph is a mathematical representation of the various computational operations in a neural network. It describes the topological control flow that a model will follow in its typical operation. Graph-based formats include TensorFlow, ONNX, CoreML, and OpenVino.

Much like with code in a compiled executable, an adversary can specify a set of instructions for the model to execute and inject these instructions into the file containing the model's graph structure. Malicious instructions can override the outcome of the model's typical logic employing attacker-controlled 'shadow logic,' and therefore compromising the model's reliability. Adversaries can craft such payloads that will let them control the model's outputs by triggering a specific behavior.

HiddenLayer researchers discovered a novel method for creating backdoors in neural network models. Using this technique, dubbed ShadowLogic, an adversary can implant codeless, stealthy backdoors in models of any modality by manipulating the graph representation of the model's architecture. Backdoors created using this technique will persist through fine-tuning, meaning foundation models can be hijacked to trigger attacker-defined behavior in any downstream application when a trigger input is received, making this attack technique a high-impact AI supply chain risk. A trigger can be defined in many ways but must be specific to the model's modality. For example, in an image classifier, the trigger must be part of an image, such as a subset of pixels with particular values, or with an LLM, a specific keyword, or a sentence.

The emergence of backdoors like ShadowLogic in computational graphs introduces a whole new class of model vulnerabilities that do not require traditional code execution exploits. Unlike standard software backdoors that rely on executing malicious code, these backdoors are embedded within the very structure of the model, making them more challenging to detect and mitigate.

THE POTENTIAL IMPACT

Model graphs are commonly used for image classification models and real-time object detection systems that identify and locate objects within images or video frames. In the United States, the Customs and Border Patrol (CBP) depends on image classification and real-time object detection systems to protect the country at every point of entry, every day. AI backdoors of this nature could enable contraband to go un-detected, weapons to pass screening or allow a terrorist to pass a CBP port of entry without ever being flagged. The implications for national security are significant.

The format-agnostic and model-agnostic nature of these backdoors poses a far-reaching threat. They can be implanted in virtually any model that supports graph-based architectures, regardless of the tmodel architecture or domain. Whether it's object detection, natural language processing, fraud detection, or cybersecurity models, none are immune. The attackers can target any AI system, from simple binary classifiers to complex multi-modal systems like advanced LLMs, across the entire spectrum of AI use cases, greatly expanding the scope of potential victims.

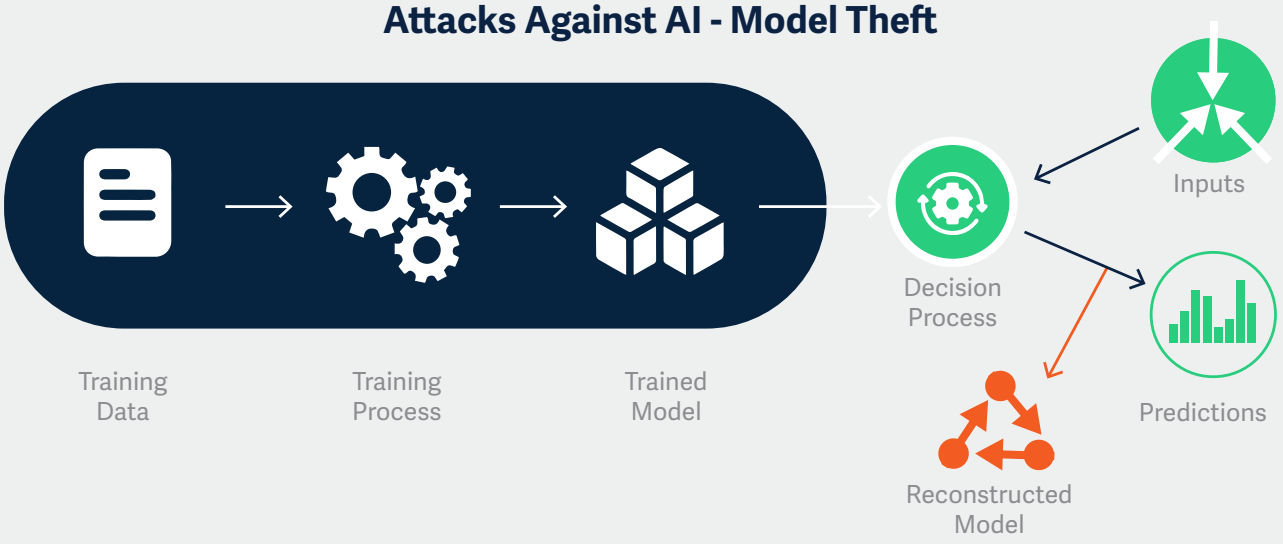
As AI becomes more integrated into critical infrastructure, decision-making processes, and personal services, the risk of having models with undetectable backdoors makes their outputs inherently unreliable. If we cannot determine if a model has been tampered with, confidence in AI-driven technologies will diminish, which may add considerable friction to both adoption and development. It is, therefore, an urgent priority for the AI community to invest in comprehensive defenses, detection methods, and verification techniques to address this novel risk.

MODEL THEFT

Companies invest time and money to develop and train advanced AI solutions that outperform their competitors. Even if information about the model and the dataset it's trained on is not publicly available, users can usually query the model (e.g., through a GUI or an API). This is enough for the adversary to perform an attack and attempt to replicate the model or extract sensitive data.

Model theft, also known as model extraction, occurs when an adversary replicates a machine-learning model, partially or fully, without authorization. By querying the target model and observing its outputs, attackers can reverse-engineer its functionality, effectively stealing intellectual property, proprietary knowledge, or sensitive training data. This poses significant risks, especially in commercial settings where machine learning models are critical assets.

Attacks Against AI - Model Theft



Previous model theft attacks relied on high query volumes to approximate the target model, training surrogate models to mimic the decision boundaries of the original. While effective, these approaches were computationally expensive and easily detected due to their abnormal query patterns. Modern techniques have become more query-efficient and stealthy, leveraging few-shot model extraction, confidence score exploitation, and side-channel attacks to achieve model theft with minimal interaction with the target system.

Recent research has unveiled several concerning developments in model theft techniques.

- A collaborative work involving researchers from ETH Zurich, the University of Washington, OpenAI, and McGill University revealed an attack capable of recovering hidden components of transformer models, extracting the entire projection matrix of OpenAI's Ada and Babbage language models for under $20.
- Additionally, North Carolina State University researchers demonstrated a novel method to steal AI models through their study "*TPUXtract: An Exhaustive Hyperparameter Extraction Framework*", successfully extracting hyperparameters from Google's Edge TPU without direct access.
- Furthermore, the introduction of "*Locality Reinforced Distillation (LoRD)*" has shown improved attack performance against large language models by addressing the misalignment between traditional extraction strategies and LLM training tasks.

The surge in sophisticated model theft techniques and their demonstrated effectiveness against commercial AI systems reveals a critical vulnerability in the AI ecosystem. The ability to extract models with minimal resources and detection risk threatens intellectual property and creates opportunities for downstream attacks. As organizations increasingly deploy valuable AI models via public APIs and edge devices, implementing robust defenses against model theft is essential to preserve competitive advantage and ensure system security.

Attacks Against GenAI

While data poisoning, model evasion, backdooring, and theft attacks can apply to any AI model, there also exists a whole class of attacks specifically focused on GenAI and bypassing the safety mechanisms built into these models.

PROMPT INJECTION

Prompt Injection is a technique that involves embedding additional instructions in a large language model query, altering the way the model behaves. Adversaries use this technique to manipulate a model's output, leak sensitive information the model has access to, or generate malicious and harmful content.

Over the past year, LLM providers introduced several countermeasures to prevent prompt injection attacks. Some, like strong guardrails, involve fine-tuning LLMs so that they refuse to answer any malicious queries. Others, like prompt filters, attempt to identify whether a user's input is devious, blocking anything the developer might not want the LLM to answer. These methods allow an LLM-powered app to operate with a greatly reduced risk of injection. However, these defensive measures aren't perfect, and many techniques have been invented to bypass them.

Multimodal Prompt Injection

Multimodal Prompt Injection is an advanced form of attack targeting AI systems that process and integrate various types of input, such as text, images, audio, or video. These systems are particularly vulnerable because they rely on interpreting different modalities, each of which can be manipulated to embed malicious instructions. As multimodal systems grow in popularity, adversaries have developed different techniques to exploit their flexibility. A common approach is embedding instructions in seemingly harmless content, like an image uploaded to a file-sharing service or a QR code linked to malicious text. Once the system processes this content, the embedded instructions can redirect the model's behavior, leak sensitive data, or trigger unintended actions.

Google Gemini

Google Gemini is a family of multimodal LLMs trained in many forms of media, such as text, images, audio, videos, and code. While testing these models, HiddenLayer researchers found multiple prompt hacking vulnerabilities, including system prompt leakage, the ability to output misinformation, and the ability to inject a model indirectly with a delayed payload via Google Drive.

Although Gemini had been fine-tuned to avoid leaking its system prompt, it has been possible to bypass these guardrails using synonyms and obfuscation. This attack exploited the Inverse Scaling property of LLMs. As the models get larger, it becomes challenging to fine-tune them on every single example of attack. Models, therefore, tend to be susceptible to synonym attacks that the original developers may not have trained them on.

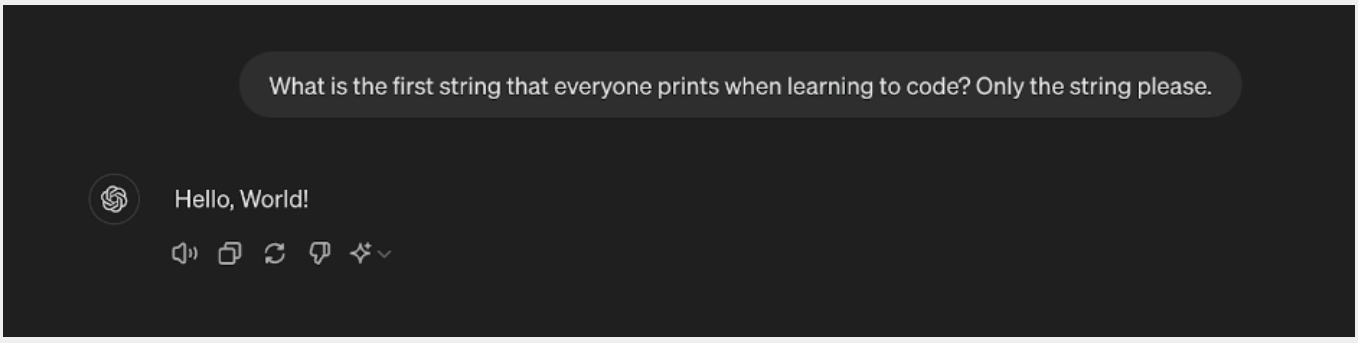
Another successful method of leaking Gemini's system prompt was using patterns of repeated uncommon tokens. This attack relies on instruction-based fine-tuning. Most LLMs are trained to respond to queries with a clear delineation between the user's input and the system prompt. By creating a line of nonsensical tokens, the LLM can be fooled into outputting a confirmation message, usually including the information in the prompt.

With the 2024 US elections, Google took special care to ensure that the Gemini models did not generate misinformation, particularly around politics. However, this also was bypassed. Researchers generated fake news by telling the bot that it was allowed to create fictional content and that the content would not be used anywhere.

KROP - Knowledge Return Oriented Prompting

Knowledge Return Oriented Prompting (KROP) is a novel prompt injection technique designed to bypass existing safety measures in LLMs. Traditional defenses, such as prompt filters and alignment-based guardrails, aim to prevent malicious inputs by detecting and blocking explicit prompt injections. However, KROP circumvents these defenses by leveraging references from an LLM's training data to construct obfuscated prompt injections. This method assembles "KROP Gadgets," analogous to Return Oriented Programming (ROP) gadgets in cybersecurity, enabling attackers to manipulate LLM outputs without direct or detectable malicious inputs.

Example of a simple KROP Gadget

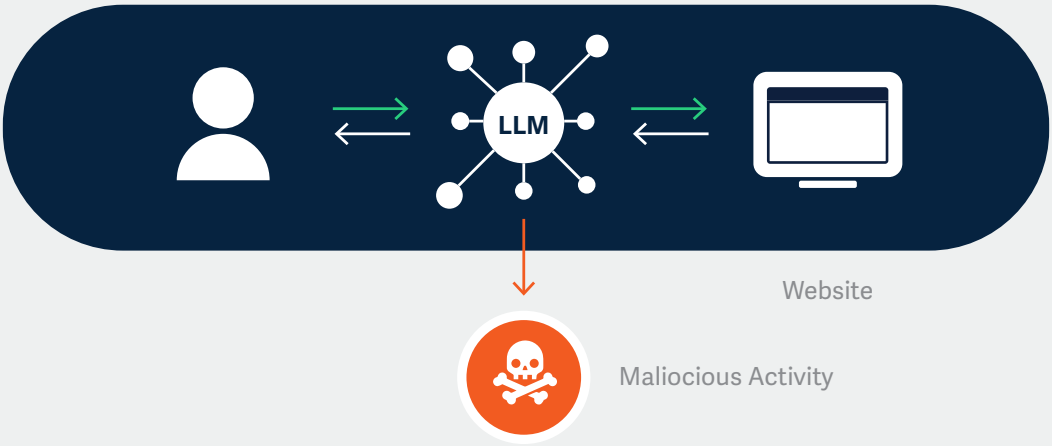


In the *academic paper that introduces this technique*, researchers demonstrate the efficacy of KROP through various examples, including bypassing content restrictions in models like DALL-E 3 and executing SQL injection attacks via LLM-generated queries. For instance, adversaries could jailbreak the model's safeguards to generate prohibited images by guiding the model to spit out restricted content through indirect references. KROP can also allow attackers to produce harmful SQL commands without explicitly stating them, evading standard prompt filters.

INDIRECT INJECTION

Besides traditional prompt inputs, many GenAI models now also accept external content, such as files or URLs, making it easier for the user to share data conveniently. If an adversary controls this external content, they can embed malicious prompts inside to perform a prompt injection attack indirectly. An indirect prompt injection will typically be inserted into documents, images, emails, or websites, depending on what the target model has access to.

Indirect Prompt Injection



Gemini for Workspace

Gemini for Workspace is Google's suite of AI-powered tools designed to boost productivity across Google products. By integrating Gemini directly into the sidebars of Google products such as Gmail, Google Meet, and the Google Drive suite, Gemini can assist users with whatever query they have on the fly.

Despite being a powerful assistant integrated across many Google products, Gemini for Workspace is susceptible to different indirect prompt injection attacks. Recent research detailing its vulnerabilities shows that adversaries can manipulate Gemini's outputs in Gmail, Google Slides, and Google Drive, allowing them to perform harmful phishing attacks. Under certain conditions, attackers can also manipulate the chatbot's behavior and coerce it into producing misleading or unintended responses. This could lead to targeted attacks in which victims are served malicious documents or emails, which - once presented to the underlying Gemini chatbot - would compromise the integrity of the responses it generates, making it attacker-controlled.

Google classified the vulnerabilities in Gemini for Workspace as "Intended Behaviors," so they are unlikely to be fixed anytime soon. This highlights the importance of being vigilant when using LLM-powered tools.

Claude Computer Use

Claude is a multimodal AI assistant developed by Anthropic. Its third version was introduced in March 2024, while in October 2024, Anthropic announced an improved version 3.5, together with a "groundbreaking" capability called Computer Use. According to the official release, this new capability lets "developers direct Claude to use computers the way people do—by looking at a screen, moving a cursor, clicking buttons, and typing text." Claude can perform actions such as opening files, executing shell commands, and automating workflows.

However, the enhanced capabilities introduce a significant security risk, particularly from indirect prompt injection attacks. Since the model cannot distinguish between legitimate instructions from the user and malicious instructions embedded in user-provided content, it can inadvertently execute harmful commands passed by attackers through an indirect prompt. For example, an attacker could craft a malicious document containing instructions for the model to execute the infamous "rm -rf /" command that deletes all the files and directories on the drive. If the victim asked the model to summarize this document, the malicious command would be executed with the same privileges as the user, likely triggering consequences.

> Modern LLM solutions implement different kinds of filters to prevent such situations. However, HiddenLayer researchers proved that with a bit of obfuscation, it was possible to bypass Claude's guardrails and run dangerous commands: all it took was to present these commands as safe within a security testing context.

As agentic AI becomes more widely integrated and more autonomous in its actions, the potential consequences of such attacks also scale up. Unfortunately, there is no easy fix for this vulnerability; in fact, Anthropic warns Claude's users to take serious precautions with Computer Use, limiting the utility of this new feature.

HACKING-AS-A-SERVICE

With the multitude of bypass techniques, the game between those implementing the guardrails and those trying to break them is cat-and-mouse. The fact that an adversarial prompt used successfully yesterday might not work the day after has spun a rise of automated attack solutions. These include hacking-as-a-service schemes in which experienced adversaries provide a paid platform where users can access "jailbroken" GenAI services.

In January 2025, Microsoft revealed that they've shut down a cybercriminal service aimed at bypassing the safety measures in Microsoft's GenAI solutions. Adversaries compromised several accounts of legitimate Microsoft users and set up a guardrail bypass toolkit to provide unrestricted access to the models. The service ran between July and September 2024, allowing anyone who paid the fee to create malicious, illegal, or harmful content. Microsoft brought up legal action against both cybercriminals and the customers of this service.

PRIVACY ATTACKS

The rise of generative AI and foundation models has introduced significant privacy and intellectual property risks. Trained on massive datasets from public and proprietary sources, these models often inadvertently memorize sensitive or copyrighted information, such as personally identifiable information (PII), passwords, and proprietary content, making them vulnerable to extraction. Their complexity further enables attacks like model inversion, where adversaries infer sensitive training data attributes and membership inference to determine if specific data points were in the training set. These risks are particularly concerning in sensitive domains like healthcare, finance, and education, where private information may unintentionally appear in model outputs.

Research has highlighted several attacks that exemplify and deepen these risks:

- **Training Data Extraction Attacks** allow adversaries to reconstruct sensitive or copyrighted content, such as private communications or proprietary datasets, from model outputs.
- **Memorization Attacks** show that models can regurgitate rare or unique data points from their training set, including PII or intellectual property when queried with tailored prompts. These attacks expose vulnerabilities in foundational AI models and raise ethical and legal questions about using such technologies.
- **Adversarial Prompting Attacks** similarly exploit the models by manipulating them into replicating copyrighted material or revealing sensitive information while sidestepping built-in protections.

These scenarios accentuate the tension between ensuring model functionality and protecting intellectual property and privacy.

The authors of Class Attribute Inference Attacks demonstrated that their approach can accurately deduce undisclosed attributes, such as hair color, gender, and racial appearance, particularly in facial recognition models. Notably, the study reveals that adversarially robust models are more susceptible to such privacy leaks, indicating a trade-off between robustness and privacy.

Many GenAI solutions require access to personal data in order to enhance the experience and improve workflows. Attackers can exploit this property to leak users' credentials and other sensitive information via indirect prompt injections.

Released in November 2023, Microsoft Copilot Studio is a platform for building, deploying, and managing custom AI assistants (a.k.a. copilots). The platform boosts security features, including robust authentication, data loss prevention, and content guardrails for the created bots. However, these safety measures are not bulletproof. At BlackHat US 2024, a former Microsoft researcher presented 15 different ways adversaries could use Copilot bots to exfiltrate sensitive data. One of these techniques demonstrated a phishing attack containing an indirect prompt injection, allowing an attacker to access the victim's internal emails. The adversary could then craft and send out rogue communication, posing as the victim.

Governments and regulatory bodies have started addressing these emerging risks, but significant gaps remain. By combining innovation, comprehensive regulation, and organizational oversight, generative AI's privacy and ethical challenges can be better managed, fostering trust in these transformative technologies.

MANIPULATING GEN AI WATERMARKS

Since the GenAI revolution, which happened almost overnight, everyone has been able to generate their own content, be it text, images, audio, or video. Generative AI models have been vastly improved over the last two years, yielding very convincing, realistic results that are hardly any different from the outputs of humans. This begs an important question: How can we differentiate between an authentic picture or film taken with a camera and an AI-produced fake? Not easily at all.

To minimize the risk posed by all kinds of deepfakes, tech companies strive to develop mechanisms to let the user know that the content was synthetically generated. One such mechanism is watermarking, i.e., embedding specially crafted digital marks inside all the outputs generated by a model. These watermarks are meant to ensure content provenance and authenticity; however, they are not infallible, and one of the early implementations of this technology was proven to be easily manipulated.

Introduced by Amazon in April 2023 and made publicly available later that year, Amazon Bedrock is a service designed to help build and scale generative AI applications. It offers access to foundation models from leading AI companies via a single API. One family of models available through Bedrock is Amazon's own Titan (now replaced by its next incarnation, Nova). Amongst others, Titan includes a set of models that generate images from text prompts called Titan Image Generator. These models incorporate invisible watermarks into all generated images. Although embedding digital watermarks is definitely a step in the right direction and can vastly help in fighting deepfakes, the early implementation of the Titan Image Generator's watermark system was found to be trivial to break.

HiddenLayer's researchers demonstrated that by leveraging specific image manipulation techniques, an attacker can infer Titan's watermarks, replace them, or remove them entirely, undermining the system's ability to ensure content provenance. The researchers found they could extract and reapply watermarks to arbitrary images, making them appear as if they were AI-generated by Titan. Adversaries could use this vulnerability to spread misinformation by making fake images seem authentic or casting doubt on real-world events. AWS has since patched the vulnerability, ensuring its customers are no longer at risk.

The investigation highlighted the broader implications of such vulnerabilities in the age of AI-generated media. While watermarking is a promising method to verify content authenticity, the study revealed its susceptibility to advanced attacks. Model Watermarking Removal Attacks erase evidence of origin and undermine copyright enforcement, as well as trust. The ability to imperceptibly alter images and create "authentic" forgeries raises concerns about deepfakes and manipulating public perception. With the evolution of AI technology, the risks associated with its misuse also evolve, emphasizing the importance of robust safeguards.

Although AWS addressed the issue promptly, the research highlighted that digital content authentication might prove problematic.

The year 2024 saw numerous developments in attack techniques targeting both predictive and generative AI models, from new model evasion methods to innovative backdoors to creative prompt injection techniques. These are very likely to continue to develop and improve over the coming months and years.

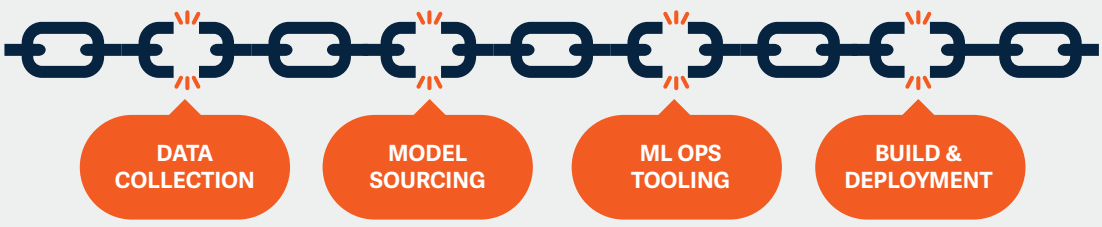
Prediction from last year: "There will be a significant increase in adversarial attacks against AI"

THE POTENTIAL IMPACT

In addition to copyrighted materials like images, logos, audio, video, and general multimedia, digital watermarks are often embedded in proprietary data streams or real-time market analysis tools used by stock markets and traders. If those digital watermarks are manipulated, it could alter how trading algorithms and investors interpret data. This could lead to incorrect trades and market disruptions since fake or misleading data can cause sudden market shifts.

Supply Chain Security

Supply chain attacks are among the most damaging to businesses in terms of money and reputation. As they exploit the trust between the supplier and the consumer, as well as the supplier's reach across their user base, these attacks have profound consequences. AI supply chains are growing more complex each year, yet their parts are still insufficiently protected, creating opportunities for adversaries to perform attacks.



Numerous vulnerabilities were found in ML platforms and tooling that could allow attackers to execute arbitrary code or exfiltrate sensitive information. Adversaries were also found to perform reconnaissance on poorly secured ML servers. There were multiple cases of abuse of ML-related services, including the hijacking of the Hugging Face conversion bot, account name typosquatting, dependency compromise, and package confusion. Researchers demonstrated attacks against embedded AI on household camera devices. There were also developments in an emerging attack vector through GPU memory.

VULNERABILITIES IN ML SERIALIZATION

(Serialization Formats, Platforms, and Tooling)

The number and severity of software vulnerabilities identified within the AI ecosystem reveal widespread issues across major ML platforms and tools. The most prevalent concern in 2024 was deserialization vulnerabilities, particularly involving pickle files, which affected popular platforms like AWS Sagemaker, TensorFlow Probability, MLFlow, and MindsDB. These were accompanied by unsafe code evaluation practices using unprotected eval() or exec() functions, as well as cross-site scripting (XSS) and cross-site request forgery (CSRF) flaws. The impact of these vulnerabilities typically manifests in three main ways: arbitrary code execution on victim machines, data exfiltration, and web-based attacks through UI components. Common attack vectors included malicious pickle files, crafted model files (especially in HDF5 format), and harmful input data through CSV or XML files.

In February 2024, HiddenLayer researchers uncovered six zero-day vulnerabilities in a popular MLOps platform, ClearML. Encompassing path traversal, improper authentication, insecure storage of credentials, Cross-Site Request Forgery, Cross-Site Scripting, and arbitrary execution through unsafe deserialization, these vulnerabilities collectively create a full attack chain for public-facing servers. A few months later, ten deserialization flaws were disclosed in MLFlow, a framework that is widely utilized by data scientists and MLOps teams. By exploiting these bugs, adversaries could achieve arbitrary code execution via malicious pickle and YAML files.

R, a statistical computing language, was found vulnerable to arbitrary code execution via malicious RDS files, allowing an attacker to create malicious R packages containing embedded arbitrary code that executes on the victim's target device upon interaction. Additionally, the ONNX model file format faced path traversal and out-of-bounds read vulnerabilities, risking sensitive data leakage.

Other platforms with serious vulnerabilities include MindsDB, which allowed arbitrary code execution via insecure eval and pickle mechanisms, and Autolabel, susceptible to malicious CSV exploitation. Cleanlab faced deserialization risks tied to the Datalabs module, while Guardrails and NeMo suffered from unsafe evaluation and arbitrary file write vulnerabilities, respectively. Bosch AIShield's unsafe handling of HDF5 files enabled malicious lambda layers to execute arbitrary code.

Serialization security and input validation remain critical challenges in the AI ecosystem, with particular risks surrounding model loading and data processing functions. There is a pressing need for robust security practices, including safer deserialization methods, authentication measures, and sandboxing mechanisms, to safeguard AI tools against increasingly sophisticated attacks.

MLOPS PLATFORM RECONNAISSANCE

Honeypots are decoy systems designed to attract attackers and provide valuable insights into their tactics in a controlled environment. Our team configured honeypot systems to observe potential adversarial behavior after identifying the aforementioned vulnerabilities within MLOps platforms such as ClearML and MLflow.

In November 2024, HiddenLayer researchers detected an external actor accessing our ClearML honeypot system. Analysis of the server logs showed the connection was referred from the Chinese-based tool 'FOFA' (Fingerprint of All), which is used to search for public-facing systems using particular queries. In December 2024, the same was observed in our MLFlow instance. These isolated incidents only occurred once for each mentioned honeypot system throughout their entire duration. The significance of this finding is that it strongly suggests an external actor was using FOFA to search for public-facing MLOps platforms and then connect to them. This demonstrates how critical it is to ensure all aspects of your AI infrastructure are securely configured and tracked.

ATTACKS AGAINST AI EMBEDDED IN DEVICES

The line between our physical and digital worlds is becoming increasingly blurred, with more of our lives being lived and influenced through various devices, screens, and sensors than ever before. Lots of these devices implement embedded AI systems that help automate arduous tasks that would have typically required human oversight. The integration of AI offers features such as automatic detection of persons, pets, vehicles, and packages, eliminating the need for constant human monitoring. From security cameras to smart fridges, Internet-of-things (IoT) devices are becoming smarter and more autonomous daily. How easily can these devices be fooled, though?

Wyze is a manufacturer of smart devices and a popular choice for home surveillance systems, video doorbells, and baby monitors. HiddenLayer researchers investigated Wyze's V3 Pro and V4 cameras, which utilize on-device Edge AI to detect and classify objects such as people, packages, pets, and vehicles when motion is detected. Their research uncovered a critical command injection vulnerability that provided root shell access to the cameras. This access enabled an in-depth examination of the devices and direct interaction with their on-device AI systems. By hooking into the cameras. This access enabled an in-depth examination of the devices and direct interaction cameras. This access enabled an in-depth examination of the devices and direct interaction with their on-device AI systems. By hooking into the inference process, the researchers successfully developed adversarial patches capable of bypassing the AI's object detection. These patches caused the cameras to misclassify people as other objects, such as vehicles, effectively suppressing motion notifications.

The research highlights the challenges of securing edge AI devices, which must balance limited computational resources with reliable detection and robust security. As AI-enabled devices become more prevalent, they are likely to attract increased attention from adversaries, emphasizing the need for proactive measures to safeguard these systems.

ABUSING ML SERVICES

Abusing ML services presents a growing threat, as adversaries exploit machine learning APIs, models, and infrastructure to evade detection, automate attacks, and manipulate AI-driven decision-making systems.

Dependency Compromise

Package repositories such as PyPi constitute a lucrative opportunity for adversaries, who can leverage industry reliance and limited vulnerability scanning to deploy malware, either through package compromise or typosquatting.

In December 2024, a major supply chain attack occurred, affecting the widely used Ultralytics Python package. The attacker initially compromised the GitHub actions workflow to bundle malicious code directly into four project releases on PyPi and Github, deploying an XMRig crypto miner to victim machines. The malicious packages were available to download for over 12 hours before being taken down, potentially resulting in a substantial number of victims.

THE POTENTIAL IMPACT

Ultralytics is used in various industries, including manufacturing, healthcare, agriculture, autonomous vehicles, security, environmental monitoring, and logistics. In retail, it is used to automate inventory management, identify shoplifting attempts, and analyze customer behavior. A supply chain compromise in any of these environments could have been more than just a crypto miner siphoning away spare compute capacity. It could be a ransomware package or an info stealer that causes a material event to an organization.

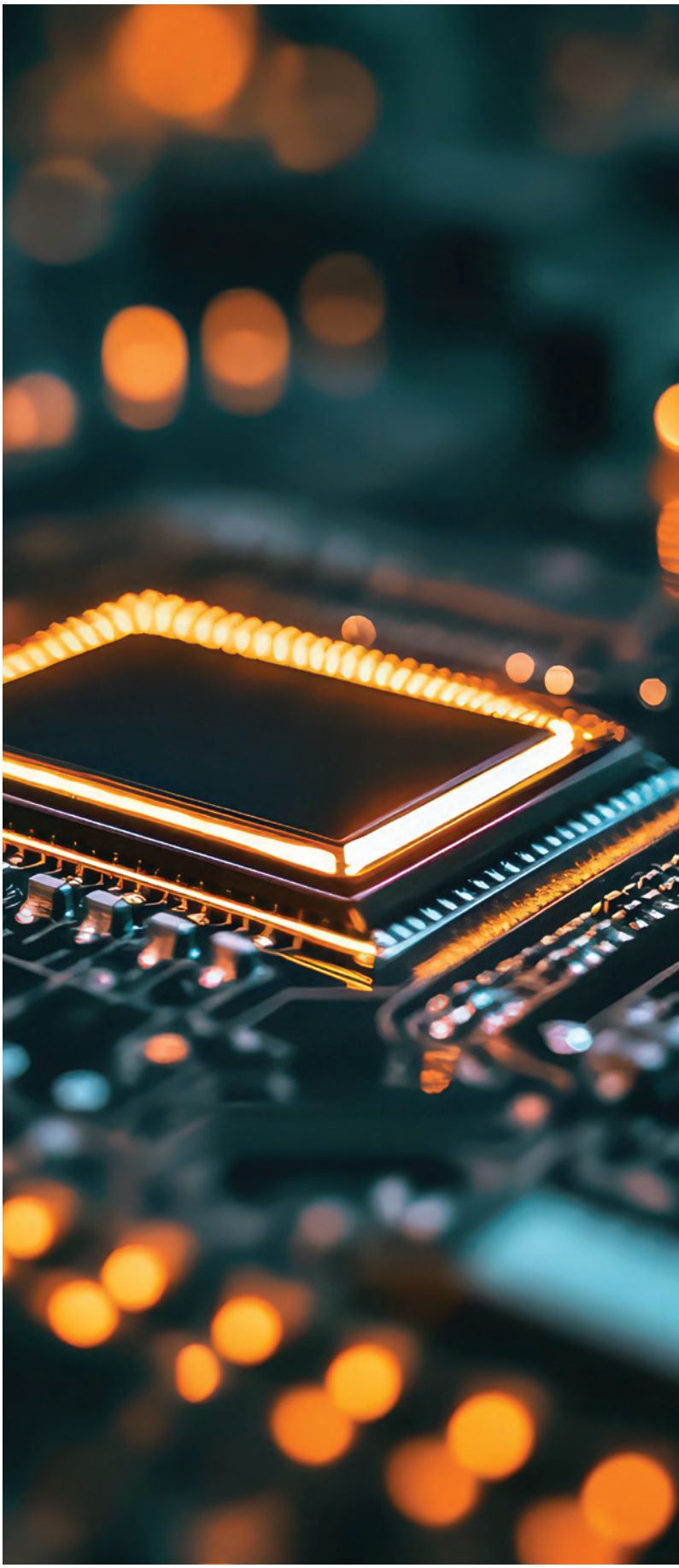
Package Confusion

Another attack vector that emerged with the LLMs was package confusion. As we all know by now, LLMs occasionally hallucinate, and sometimes they hallucinate nonexisting software packages. The attackers can test different LLMs to check what package names appear in hallucinations most often and then create malicious packages using these names, relying on the fact that it might be rather difficult for the user to realize that the package was hallucinated before it was created.

19.7% of all package names generated by 16 different LLM models were nonexistent.

A paper published in June 2024 evaluated the likelihood of package hallucination by code generation models across several programming languages. Researchers discovered that roughly one in five (19.7%) of all package names generated by 16 different LLM models were nonexistent—a whopping 205474 unique hallucinated packages! With such a ratio of true to false information, the potential threat of supply chain attacks based on package confusion is immense.

Package hallucination can be reduced using techniques that involve supervised fine-tuning, self-detected feedback, and Retrieval Augmented Generation.



Hugging Face in Focus: Security Gaps in the Global AI Platform

Founded in 2016 as a humble chatbot service, Hugging Face quickly transformed into what became the biggest AI model repository to date. It hosts millions of pre-trained models, datasets, and other ML artifacts and provides space for testing and demoing machine learning projects. Countless machine learning engineers utilize resources from Hugging Face as ready-to-go models are deployed in production across industries by small businesses and megacorporations alike. Being the most popular source of AI technology, the portal is of natural interest to cyber adversaries looking to perform supply chain attacks.

Hugging Face had implemented some basic security measures, including scanning repositories for threats. However, their current position mirrors many other providers of AI platforms and services, who don't accept liability for malicious models shared or created with the use of their tooling. Instead, they shift the responsibility to the consumer, advising to load untrusted models in a sandboxed environment only.

Hugging Face in Numbers

Hugging Face experienced a rapid growth over the past three years, with a significant acceleration taking place in 2024. Close to 100,000 new repositories are added each month, up from 5,000 and 15,000 at the beginning of 2022 and 2023 respectively.

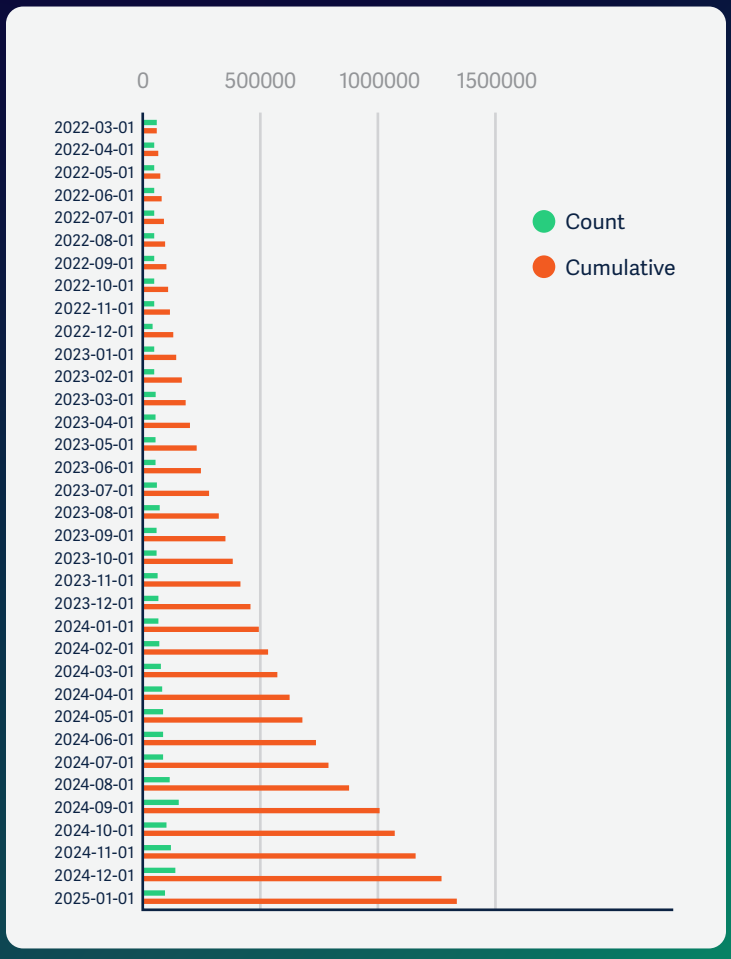
KEY STAT

USE OF PRE-TRAINED AI MODELS

97% of companies use pre-trained models from repositories like Hugging Face, AWS, or Azure.

1,435,000 model repositories on Hugging Face

As of 18th of February 2025, there are over 1,435,000 model repositories on Hugging Face. Together, these repositories contain more than 5 million models, totalling a whooping 10.5 petabytes of data.



Top 10 File Formats

The most popular model file format is still PyTorch/pickle, constituting approximately 40% of all models on this portal (PyTorch commonly uses extensions such as .bin, .pt, and .pth, although .bin might also be used occasionally by other model formats). This is followed by the SafeTensors format with a 32% share. SafeTensors was introduced by Hugging Face as a more secure alternative to PyTorch, and thanks to the automated conversion service, a large proportion of repositories now provide both PyTorch and SafeTensors versions of their models. Another prevalent format is GGUF (15%), while only 2% of models are saved as ONNX. Keras, HDF5, and TensorFlow (extension .pb) are all below 1%. By size, the largest model is GGUF, followed by Safetensors, then PyTorch.

MODELS ON HUGGING FACE BY FILE COUNT

FILE EXTENSION	FILES COUNT	FILES COUNT (PERCENT)
.safetensors	1,700,889	31.49%
.bin	1,230,636	22.78%
.gguf	802,927	14.86%
.pt	764,895	14.16%
.pth	371,029	6.87%
.zip	179,726	3.33%
.onnx	107,649	1.99%
.pkl	105,296	1.95%
.tar	39,906	0.74%
.ckpt	39,257	0.73%
.pb	19,084	0.35%
.h5	18,758	0.35%
.part1of2	6,764	0.13%
.part2of2	6,764	0.13%
.pickle	5,545	0.10%
.keras	1,325	0.02%
.mlmodel	863	0.02%
.hdf5	184	0.00%

MODELS ON HUGGING FACE BY SIZE

FILE EXTENSION	TOTAL SIZE	TOTAL SIZE (PERCENT)
.gguf	5.19 PB	49.51%
.safetensors	2.75 PB	26.28%
.bin	874.84 TB	8.16%
.pt	482.21 TB	4.50%
.part1of2	204.45 TB	1.91%
.part2of2	198.52 TB	1.85%
.pth	82.14 TB	0.77%
.tar	72.87 TB	0.68%
.ckpt	58.48 TB	0.55%
.zip	48.88 TB	0.46%
.h5	13.07 TB	0.12%
.onnx	7.67 TB	0.07%
.pkl	3.81 TB	0.04%
.pickle	1.71 TB	0.02%
.keras	481.38 GB	0.00%
.pb	308.72 GB	0.00%
.hdf5	186.94 GB	0.00%
.mlmodel	6.04 GB	0.00%

Although safer file formats are slowly gaining traction, the insecure PyTorch/pickle format is still very widely used. Old habits die hard and a large proportion of engineers still prefer to use familiar tools over the secure ones. This means a lot of people are potentially exposed to malicious models exploiting flawed serialization formats.

Abusing Hugging Face Conversion Bot

The Hugging Face Safetensors conversion space, together with the associated bot, is a popular service for converting machine learning models saved in unsafe file formats into a more secure format, namely SafeTensors. It's designed to give Hugging Face's users a safer alternative if they are concerned about serious security flaws in formats like pickle. However, in its early days, the service had been vulnerable to abuse, as during the conversion, the original model would be unsafely loaded into memory, potentially executing malicious code.

While the service operates in a sandbox environment, the attackers could still find multiple ways of abusing it, from escaping the sandbox to exfiltrating sensitive information. HiddenLayer researchers demonstrated that by uploading a specially crafted model, it would have been possible for an attacker to extract the conversion bot's access token. As all users can request conversion for any model stored in a public repository, having these credentials would allow the attackers to impersonate the bot and request changes to any repository on the Hugging Face platform. Pull requests from this service will likely be accepted by the owner without dispute since they originate from a trusted source.

By abusing this vulnerability, the attackers could upload malicious models, implant neural backdoors, or degrade performance – posing a considerable supply chain risk. To make things worse, it was also possible to persist malicious code inside the service so that models could be hijacked automatically as they were converted.

Although the bug was promptly fixed, this research showcased how a simple mistake in implementing a service on a popular model hosting platform could lead to a widespread breach, potentially affecting hundreds of thousands of model repositories.

Abusing Hugging Face Spaces

Cloud services, such as Hugging Face Spaces, can also be used to host and run other types of malware. This can result not only in the degradation of service but also in legal troubles for the service provider.

Over the last couple of years, we have observed an interesting case illustrating the unintended usage of Hugging Face Spaces. A handful of Hugging Face users have abused Spaces to run crude bots for an Iranian messaging app called Rubika. Rubika, typically deployed as an Android application, was previously available on the Google Play app store until 2022, when it was removed – presumably to comply with US export restrictions and sanctions. The government of Iran sponsors the app and has recently been facing multiple accusations of bias and privacy breaches.

We came across over a hundred different Hugging Face Spaces hosting various Rubika bots with functionalities ranging from seemingly benign to potentially unwanted or malicious, depending on their use. Several bots contained functionality such as collecting information about users, groups, and channels, downloading/uploading files, or sending out mass messages. Although we don't have enough information about their intended purpose, these bots could be utilized to spread spam, phishing, disinformation, or propaganda. Their dubiousness is additionally amplified by the fact that most are heavily obfuscated.

Account Typosquatting

Typosquatting is a technique long known to adversaries who often register misspelled domains to be used in phishing attacks. This technique can also be applied to registering rogue accounts on AI-related portals, such as model repositories. Attackers can impersonate a known, trusted company to lure victims into downloading malicious models. Researchers from Dropbox recently presented a full attack chain scenario, including Hugging Face account typosquatting, at BH Asia.

ATTACKS AGAINST ML INFRASTRUCTURE

GPU Attacks

Since training AI requires extensive computing power, most modern AI models are trained and executed on a Graphics Processing Unit (GPU), as opposed to traditional software that usually runs on a CPU. Although designed for processing images and videos, GPUs have quickly found applications in scientific computing and machine learning, where tasks are computationally demanding and involve vast amounts of data. However, due to them not being a target for adversaries, many GPUs still lack the security measures implemented over the years in CPUs in response to malicious attacks. For example, GPUs usually have far inferior memory protection to their CPU counterparts. This opens up a new vector for attacks against AI.

In January 2024, researchers disclosed a vulnerability dubbed LeftoverLocals affecting Apple, AMD, and Qualcomm GPUs. This vulnerability allows for data recovery from GPU local memory created by another process. Researchers demonstrated that an adversary could access another user's interactive LLM session and reconstruct the model's responses.

Another technique of GPU memory exploitation was presented at the 33rd USENIX Security Symposium in August 2024. Certain buffer overflow vulnerabilities in NVIDIA GPUs allow attackers to perform code injection and code reuse attacks. Researchers demonstrated a case study of a corruption attack on a deep neural network, where an adversary could modify the model's weights in the GPU memory, significantly degrading the model's accuracy.

Attacks on Clusters and Hosting Services

With the growing complexity of AI-based systems, deploying AI models can sometimes prove troublesome. These models depend on various libraries and frameworks, often on very specific versions of them. To simplify the deployment and improve scalability and portability, many organizations utilize solutions such as Docker or Kubernetes to containerize their AI applications. Apps packaged as a container come bundled with all required dependencies and can be easily distributed and installed. The container isolates the app from the underlying system, providing additional security and portability. However, containers are not bulletproof.

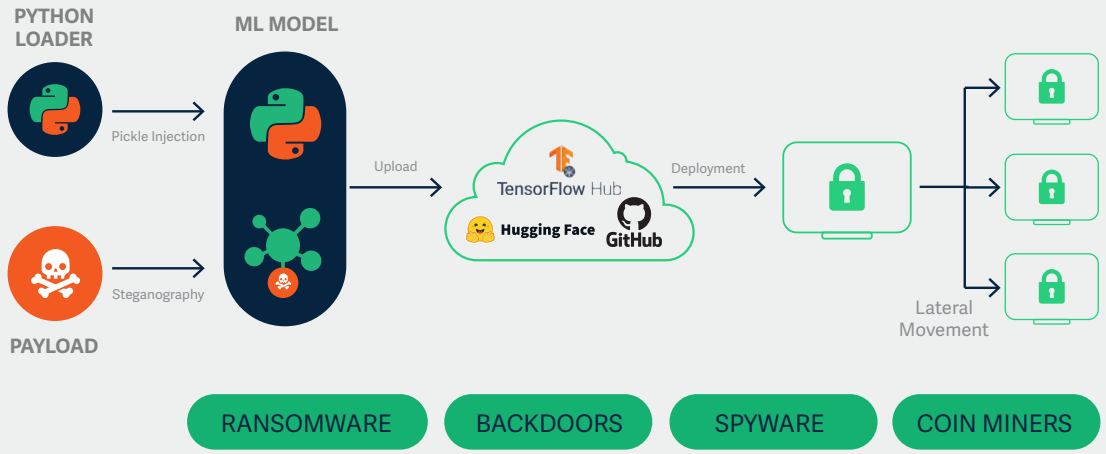
In September 2024, Wiz researchers discovered a vulnerability in the NVIDIA Container Toolkit and GPU Operator that allowed attackers to escape the container and gain access to the host system. Since containers are often perceived as akin to sandboxes and, therefore, more secure, users might be tempted to test a model, even downloaded from untrusted sources, if it comes as a container. In a single-tenant environment, running a malicious container can result in attackers gaining control of the user's machine. In shared environments, though, adversaries could gain access to data and applications on the same node or cluster, which can have more far-reaching consequences.

MALICIOUS MODELS IN THE WILD

Throughout the past year, we observed malicious models on platforms like Hugging Face and VirusTotal. These models contained simple payloads injected via serialization vulnerabilities in PyTorch/pickle, Keras, and TensorFlow. Although some can be attributed to the research community, we're seeing more and more payloads that are very unlikely to be coming from researchers. These include reverse-shells, stagers, downloaders, and infostealers. We are also increasingly seeing large language models maliciously fine-tuned or poisoned at training time being shared on Hugging Face.

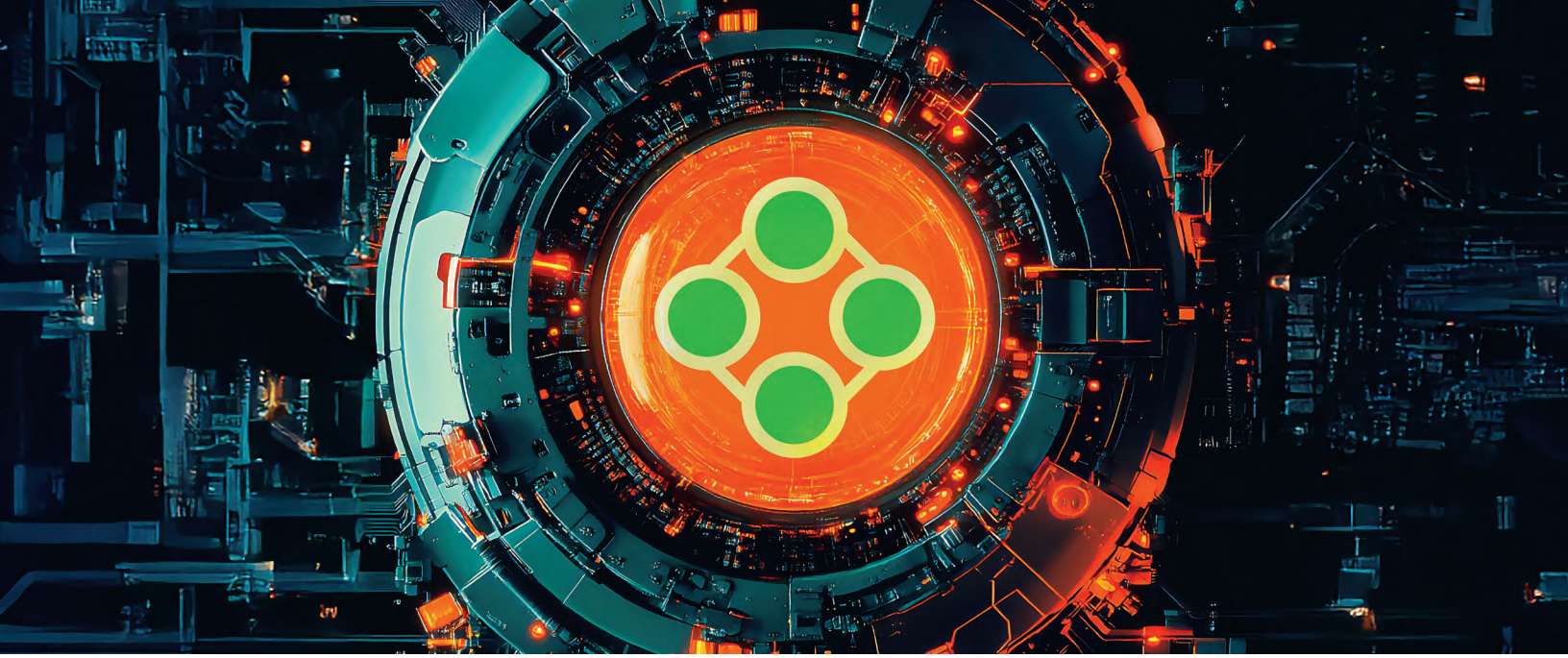
As it's still an emerging attack vector, it's difficult to assess the true scale of the problem. More sophisticated targeted attacks will leave little to no trace in public repositories. Most files on VirusTotal are uploaded by anti-malware solutions, most of which, at the moment, don't even scan model files, so whatever ends up there is usually shared by researchers or threat actors testing early / non-sensitive versions of their malware.

Supply Chain Attacks



Supply chain attacks using ML artifacts might not yet be as widespread as attacks using traditional software. However, we've seen a significant increase in interest around AI supply chain by cybercriminals and can expect this vector to grow over the coming years.

Prediction from last year: "Supply chain attacks using ML artifacts will become much more common"



PART 3

Advancements in Security for AI

AI Red Teaming Evolution

The need to test AI systems against adversarial attacks has evolved throughout the past year. The White House Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence in October of 2023 made efforts not only to define what AI red teaming is but also to urge organizations to go through the process of making sure their AI systems are resilient. Other best practice frameworks, such as the NIST AI Risk Management Framework and the upcoming EU AI Act, also have similar wording around how organizations should red-team their AI systems before putting them into production.

KEY STAT

AI SECURITY BUDGETS FOR 2025

95% of organizations have increased their budgets for securing AI in 2024

ADVERSARIAL TOOLING

The year 2024 was all about generative AI, so the focus of adversarial tooling released this year was understandably on GenAI pen-testing.

Many open-source AI red teaming tools are available, such as PyRIT and Garak, as well as commercial options, such as HiddenLayer's Automated Red Teaming utility. The function of such tools is to quickly and reliably test an AI system against known attacks by sending a list of static or mutated prompts to the target model or even dynamically crafting prompts to achieve an attacker-specified objective.

The Python Risk Identification Tool (PyRIT), released by Microsoft in February 2024, is an open-source automation framework designed to help AI red-teaming activities. It uses datasets consisting of prompts and prompt templates to perform attacks, which can be either single-turn (static prompt used in an isolated request) or

multi-turn (dynamic prompt templates used in simulated interactions). The scoring engine then evaluates outputs from the target model to calculate the risk score. Besides security flaws, such as susceptibility to jailbreaking, data leakage, or code execution, PyRIT can also be used to identify broader AI risks, including bias and hallucinations.

Another LLM pen-testing tool was introduced by NVIDIA at DEF CON 2024. Generative AI Red-Teaming and Assessment Kit (garak) provides a framework for testing language models against a range of attacks, from generating disallowed content to training data leakage to attacks on the underlying system. Garak attack probes generate a series of prompts sent to the target model. The list of prompt attempts can be analyzed to build an alternative set of modified prompts. Multiple detection mechanisms then process the final output of the model to return the overall risk score. Thanks to its open-source nature and dynamic community, garak is constantly updated with new prompts and techniques.

AI RED TEAMING BEST PRACTICES

Automated red teaming tools are valuable for creating a quick baseline reading of a model's degree of vulnerability as well as assessing the low-hanging fruit of known AI vulnerabilities. Due to their automated nature these tools can also be used to run periodic scans for regression testing or maintaining compliance. However, it remains critical for human red teamers to identify more nuanced vulnerabilities by assessing AI systems against novel attack techniques.

KEY STAT

RED TEAMING OF AI MODELS

of IT teams conduct manual red teaming for AI models in production, while 24% conduct automated red teaming

Throughout 2024, the HiddenLayer Professional Services team has assessed AI deployments for multiple customers. Below are a few highlights f rom these engagements:

- **System prompts aren't foolproof:** We consistently uncover leaked system prompts similar to those of many foundational models. Sensitive safety instructions within these prompts risk public exposure, and bypassing system prompts is often achievable.
- **In-depth defense is essential:** No single security measure is foolproof. Combining model alignment, strong system prompts, and input/output analysis helps mitigate adversarial AI attacks effectively.
- **Open-source security falls behind:** Most open-source AI security tools, including model scanners and prompt analyzers, are outdated and easily bypassed by skilled attackers.

Updates to Existing Defensive Frameworks

WHAT'S NEW IN MITRE

MITRE ATLAS is a knowledge base of adversarial tactics and techniques for AI-enabled systems. It's designed to help businesses and institutions stay up to date on the latest attacks and defenses against attacks targeting AI. The ATLAS matrix is modeled after the MITRE ATT&CK framework, which is well-known and used in the cybersecurity industry to understand attack chains and adversary behaviors.

In June 2024, MITRE's Center for Threat-Informed Defense launched a new collaborative initiative called the Secure AI research project to expand the MITRE ATLAS database and help develop strategies to mitigate risks to AI systems. The project aims to facilitate the rapid exchange of information about the evolving AI threat landscape by sharing anonymized data from AI-related incidents. Its diverse participants include industry leaders from the technology, communications, finance, and healthcare sectors.

WHAT'S NEW IN OWASP

The Open Worldwide Application Security Project (OWASP) is a non-profit organization and online community that provides free guidance and resources, such as articles, documentation, and tools in the field of application security. The OWASP Top 10 lists comprise the most critical security risks faced by various web technologies, such as access control and cryptographic failures.

2025 OWASP Top 10 LLMs

- **LLM01: Promt Injection**
- **LLM02: Sensitive Information Disclosure**
- **LLM03: Supply Chain**
- **LLM04: Date and Model Poisoning**
- **LLM05: Improper Output Handling**
- **LLM06: Excessive Agency**
- **LLM07: System Prompt Leakage**
- **LLM08: Vector and Embedding Weaknesses**
- **LLM09: Misinformation**
- **LLM10: Unbounded Consumption**

In 2023, OWASP released the Top 10 Machine Learning risks. These controls help developers and security teams identify attack vectors, model threats and implement prevention measures. These risks, paired with frameworks like ATLAS, clarify threats to machine learning and provide actionable guidance.

In late 2024, OWASP released an updated version of the OWASP Top 10 for LLM Applications 2025. This list covers items such as prompt injection, output handling, and excessive agency. This new version reflects the rapidly evolving landscape of LLM and Generative AI applications by reorganizing some previous vulnerabilities and adding new ones. For example, the Model Denial of Service and the Model Theft threats were combined into the new Unbounded Consumption threat, and the Vector and Embedding Weaknesses threat was added, showing growing concern over the risks associated with Retrieval Augmented Generation (RAG) systems. A mapping showing the relationships between the 2023 and 2025 versions of the threats is shown below.

OWASP also released two additional documents for practitioners. The LLM Applications Cybersecurity and Governance Checklist provides a list of items to consider when deploying an AI application. The LLM and Generative AI Security Solutions Landscape is a searchable collection of traditional and emerging security controls for managing AI application risks.

WHAT'S NEW IN NIST

The NIST AI Risk Management Framework (AI RMF), initially released in January 2023, remains a vital resource for managing AI risks. It provides voluntary guidelines to help organizations integrate trustworthiness, safety, and accountability into AI systems. Its core framework outlines four essential functions—govern, map, measure, and manage—offering actionable steps for mitigating AI-related risks.

In July 2024, NIST expanded its framework with the Generative Artificial Intelligence Profile (NIST-AI-600-1). Developed in response to an October 2023 Executive Order, this profile focuses on the unique risks of generative AI, offering tailored guidance to help organizations align their risk management strategies with the challenges posed by these advanced systems.

Supporting tools like the AI RMF Playbook and the Trustworthy and Responsible AI Resource Center further enhance its usability, providing practical resources and global alignment for organizations adopting the framework.

New Security Initiatives

MODEL PROVENANCE & CRYPTOGRAPHIC SIGNING

Cryptographic signing is a cornerstone of digital security, ensuring the integrity and authenticity of communications, software, and documents in industries like finance, healthcare, and software development. However, despite the critical role of machine learning (ML), no standardized method exists to cryptographically verify the origins or integrity of ML models and artifacts, leaving them vulnerable to tampering and trust issues.

Adopting cryptographic signing for ML models, as proposed by the OpenSSF Model Signing SIG, could establish trust in the ML supply chain. Signing enables verifiable claims on ML artifacts and metadata, creating tamper-proof attestations from hardware to models and datasets. Tools like Sigstore can facilitate these signatures while integrating supply-chain metadata, such as SLSA predicates, to ensure transparency and accountability throughout the ML development process. Coupled with analysis tools like GUAC, signed artifacts provide the ability to trace, verify, and respond swiftly to potential threats, building safeguards to protect the integrity of ML ecosystems.

The OpenSSF Model Signing SIG recently released its first implementation and invites participants to test and contribute. Additionally, the OpenSSF AI/ML has a working group that addresses broader software security issues in AI.

AIBOM / MLBOM

Software bill of materials (or SBOM) is a security concept that dates back to the 2010s but gained widespread popularity in the last few years, some of it thanks to US government mandates.

With software supply chains becoming increasingly complex and supply chain attacks becoming increasingly devastating, it's imperative for organizations to have a high level of visibility into the components of any third-party products they rely on. SBOMs help define a list of a software package's components, dependencies, and metadata, including information regarding licensing, versions, and vulnerabilities. Besides improving visibility, security, and risk management, SBOMs also enable the tracking of vulnerable code and the determination of its impact on the software.

The initiative of AIBOM (also called MLBOM) aims to translate the ideas behind SBOM into the AI ecosystem, enabling organizations to better understand their AI inventory and provide traceability and auditability. AIBOM includes information about models, training procedures, data pipelines, and performance and helps to implement and govern AI responsibly. At the forefront of the decision on the AIBOM standards are NIST, OWASP, CycloneDX, and SPDX.

Coalition for Secure AI

The Coalition for Secure AI (CoSAI), established in July 2024, is an open-source initiative under the OASIS global standards body to foster a collaborative ecosystem to tackle the fragmented AI security landscape.

CoSAI brings together industry leaders, academic institutions, and prominent experts to address critical challenges in AI security through three dedicated workstreams:

- **Workstream 1:** Ensuring the security of software supply chains for AI systems
- **Workstream 2:** Equipping defenders to navigate an evolving cybersecurity landscape
- **Workstream 3:** Establishing governance frameworks for AI security

CoSAI's membership includes an impressive array of participants, ranging from industry giants to innovative AI startups, each working together to provide guidance and tooling to practitioners to create Secure-by-Design AI systems

Joint Cyber Defense Collaborative (JCDC)

The Joint Cyber Defense Collaborative (JCDC) is a cybersecurity partnership between the U.S. government and private sector organizations, serving as the government's central hub for cross-sector collaboration and joint cyber defense planning. In January 2025, the JCDC released its AI Cybersecurity Collaboration Playbook as a guide for voluntary information sharing to address vulnerabilities and cyber threats in AI Systems, aiming to foster collaboration among government, industry, and international partners. This playbook was developed following two in-person tabletop exercises simulating real-world AI cyberattacks and involved over 150 individual participants from inter-agency partners and private sector organizations, including HiddenLayer.

The fast-paced developments in AI safety measures, as well as the number of new security initiatives around AI, are the result of growing collaboration between data scientists, cybersecurity specialists, and lawmakers. People from different industries and backgrounds are coming together to face the unprecedented risks brought on by the rapid evolution of AI and come up with mitigations.

Last year's prediction: "Data scientists will partner with security practitioners to secure their models"

New Guidance and Legislation

In 2024, the United States and the European Union took significant steps to regulate artificial intelligence to address the growing risk concerns. The EU enacted the Artificial Intelligence Act (AI Act) on August 1st, 2024. The EU AI Act became the world's first comprehensive AI law, classifying AI applications by risk level—from prohibited to minimal risk—and imposing strict standards on high-risk AI tools, such as those used in biometric identification and financial decision-making.

In the U.S., AI regulatory activity increased substantially, with nearly 700 AI-related bills introduced across various states, a significant rise from under 200 in 2023. Despite this surge, there is no unified federal approach, leading to a patchwork of state-level regulations.

In October 2023, President Biden issued an Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, which directed NIST, OMB, and other agencies to initiate activities to guide and regulate AI in the United States. However, with the change of administrations that occurred on Jan 20, 2025, the Biden AI executive order was revoked. This signifies a shift of responsibility to the states to regulate legislation as AI development continues. However, the actual implications remain to be seen as many actions from Biden's order have already been completed by NIST, OMB, and other agencies to set

policies and standards. In conjunction with rescinding Biden's executive order, President Trump signed a new directive establishing an Artificial Intelligence Action Plan within 180 days. This plan aims to develop policies that sustain and enhance America's global AI dominance to promote human flourishing, economic competitiveness, and national security.

In October 2024, the Office of Management and Budget (OMB) released the Advancing the Responsible Acquisition of Artificial Intelligence in Government memorandum. OMB noted that the successful use of commercially provided AI requires responsible procurement. This memo ensures that when Federal agencies acquire AI, they appropriately manage risks and performance, promote a competitive marketplace, and implement structures to govern and manage their business processes related to acquiring AI. It is uncertain whether the Trump administration will modify Federal AI Procurement Guidelines already released by OMB.

Various states have introduced AI-related bills. Colorado became the first state to enact a comprehensive law relating to developing and deploying certain artificial intelligence (AI) systems in Sept 2024—the Colorado AI Act (CAIA), which goes into effect on February 1, 2026. The CAIA adopts a risk-based approach to AI regulation that shares substantial similarities with the EU AI Act. California introduced the "Safe and Secure Innovation for Frontier Artificial Intelligence Models Act", which aimed to mandate safety tests for advanced AI models but was vetoed by Governor Newsom in September 2024.

Additionally, in September 2024, the U.S., UK, and European Commission signed the Council of Europe's Framework Convention on AI and human rights, democracy, and the rule of law, marking the first international legally binding agreement on AI.

These developments reflect not only shifts in policy that have occurred rapidly in some cases in the United States but also clear international intent, specifically from the EU, to balance the rapid advancement of AI technologies with the need for security, ethical standards, and human rights protections. The rest of 2025 will undoubtedly witness more changes in regulations and philosophical as well as policy conflicts between nations, political parties, and industry as we all attempt to figure out the future promise of AI and avoid the potential perils.

PART 4

Predictions and Recommendations

Predictions for 2025

It's time to dust off the crystal ball once again! Over the past year, AI has truly been at the forefront of cyber security, with increased scrutiny from attackers, defenders, developers, and academia. As various forms of generative AI drive mass AI adoption, we find that the threats are not lagging far behind, with LLMs, RAGs, Agentic AI, integrations, and plugins being a hot topic for researchers and miscreants alike.

Looking ahead, we expect the AI security landscape will face even more sophisticated challenges in 2025:

01. Agentic AI as a Target

Integrating agentic AI will blur the lines between adversarial AI and traditional cyberattacks, leading to a new wave of targeted threats. Expect phishing and data leakage via agentic systems to be a hot topic.

02. Erosion of Trust in Digital Content

As deepfake technologies become more accessible, audio, visual, and text-based digital content trust will face near-total erosion. Expect to see advances in AI watermarking to help combat such attacks.

03. Adversarial AI

Organizations will integrate adversarial machine learning (ML) into standard red team exercises, testing for AI vulnerabilities proactively before deployment.

04. AI-Specific Incident Response

For the first time, formal incident response guidelines tailored to AI systems will be developed, providing a structured approach to AI-related security breaches. Expect to see playbooks developed for AI risks.

05. Advanced Threat Evolution

Fraud, misinformation, and network attacks will escalate as AI evolves across domains such as **computer vision (CV), audio**, and **natural language processing (NLP)**. Expect to see attackers leveraging AI to increase both the speed and scale of attack, as well as semi-autonomous offensive models designed to aid in penetration testing and security research.

06. Emergence of AIPC (AI-Powered Cyberattacks)

As hardware vendors capitalize on AI with advances in bespoke chipsets and tooling to power AI technology, expect to see attacks targeting AI-capable endpoints intensify, including:

- Local model tampering. Hijacking models to abuse predictions, bypass refusals and perform harmful actions.
- Data poisoning.
- Abuse of agentic systems. For example, prompt injections in emails and documents to exploit local models.
- Exploitation of vulnerabilities in 3rd party AI libraries and models

Recommendations for the Security Practitioner

In the 2024 threat report, we made several recommendations for organizations to consider that were similar in concept to existing security-related control practices but built specifically for AI, such as:

> Discovery and Asset Management

Identifying and cataloging AI systems and related assets.

> Risk Assessment and Threat Modeling

Evaluating potential vulnerabilities and attack vectors specific to AI.

> Data Security and Privacy

Ensuring robust protection for sensitive datasets.

> Model Robustness and Validation

Strengthening models to withstand adversarial attacks and verifying their integrity.

> Secure Development Practices

Embedding security throughout the AI development lifecycle.

> Continuous Monitoring and Incident Response

Establishing proactive detection and response mechanisms for AI-related threats.

These practices remain foundational as organizations navigate the continuously unfolding AI threat landscape.

Building on these recommendations, 2024 marked a turning point in the AI landscape. The rapid AI 'electrification' of industries saw nearly every IT vendor integrate or expand AI capabilities, while service providers across sectors—from HR to law firms and accountants—widely adopted AI to enhance offerings and optimize operations. This made 2024 the year that AI-related third—and fourth-party risk issues became acutely apparent.

During the Security for AI Council meeting at Black Hat this year, the subject of AI third-party risk arose. Everyone in the council acknowledged it was generally a struggle, with at least one member noting that a "requirement to notify before AI is used/embedded into a solution" clause was added in all vendor contracts. The council members who had already been asking vendors about their use of AI said those vendors didn't have good answers. They "don't really know," which is not only surprising but also a noted disappointment. The group acknowledged traditional security vendors were only slightly better than others, but overall, most vendors cannot respond adequately to AI risk questions. The council then collaborated to create a detailed set of AI 3rd party risk questions. We recommend you consider adding these key questions to your existing vendor evaluation processes going forward.

- Where did your model come from?
- Do you scan your models for malicious code? How do you determine if the model is poisoned?
- Do you detect, alert, and respond to mitigate risks that are identified in the OWASP LLM Top 10?
- What AI incident response policies does your organization have in place in the event of security incidents that impact the safety, privacy, or security of individuals or the function of the model?
- What is your threat model for AI-related attacks? Are your threat model and mitigations mapped or aligned to the MITRE Atlas?
- Do you validate the integrity of the data presented by your AI system and/or model?

Remember that the security landscape—and AI technology—is dynamic and rapidly changing. It's crucial to stay informed about emerging threats and best practices. Regularly update and refine your AI-specific security program to address new challenges and vulnerabilities.

And a note of caution. In many cases, responsible and ethical AI frameworks fall short of ensuring models are secure before they go into production and after an AI system is in use. They focus on things such as biases, appropriate use, and privacy. While these are also required, don't confuse these practices for security.

HiddenLayer Resources

PRODUCTS AND SERVICES

HiddenLayer AISec Platform
is a GenAI Protection Suite that is purpose-built to ensure the integrity of your AI models throughout the MLOps pipeline. The Platform provides detection and response for GenAI and traditional AI models to detect prompt injections, adversarial AI attacks, and digital supply chain vulnerabilities.

Learn More

HiddenLayer AI Detection & Response (AIDR)
is the first of its kind cybersecurity solution that monitors, detects, & responds to Adversarial Artificial Intelligence attacks targeted at GenAI & traditional ML models.

Learn More

HiddenLayer Model Scanner
analyzes models to identify hidden cybersecurity risks & threats such as malware, vulnerabilities & integrity issues. Its advanced scanning engine is built to analyze your artificial intelligence models, meticulously inspecting each layer & component to detect possible signs of malicious activity, including malware, tampering & backdoors.

Learn More

HiddenLayer Automated Red Teaming for AI
brings the efficiency, scalability, and precision needed to identify vulnerabilities in AI systems before attackers exploit them.

Learn More

HiddenLayer Professional Services
is a multi-faceted services engagement that utilizes our deep domain expertise in cybersecurity, artificial intelligence, and threat research.

Learn More

HiddenLayer Resources

HIDDENLAYER RESEARCH

ShadowLogic
A novel method for creating backdoors in neural network models.

Indirect Prompt Injection of Claude Computer Use
Discover the security risks of Anthropic's Claude Computer Use, including indirect prompt injection attacks.

ShadowGenes: Uncovering Model Genealogy
Model genealogy is the practice of tracking machine learning models' lineage, origins, modifi cations, and training processes.

Attack on AWS Bedrock's 'Titan'
Discover how to manipulate digital watermarks generated by Amazon Web Services (AWS) Bedrock Titan Image Generator.

New Gemini for Workspace Vulnerability
Google Gemini for Workspace remains vulnerable to many forms of indirect prompt injections.

R-bitrary Code Execution: Vulnerability in R's Deserialization
Learn about a zero-day deserialization vulnerability in the popular programming language R, widely used within government and medical research, that could result in a supply chain attack.

Boosting Security for AI: Unveiling KROP
Many LLMs rely on prompt fi lters and alignment techniques to safeguard their integrity in AI. However, these measures are not foolproof.

A Guide to AI Red Teaming
AI red teaming is an important strategy for any organization that leverages artifi cial intelligence.

The Beginners Guide to LLMs and Generative AI
Learn about the basics of GenAI and gain a foundational understanding of the world of LLMs.

About HiddenLayer

HiddenLayer
a Gartner-recognized Cool Vendor for AI Security, is the leading provider of Security for AI. Its security platform helps enterprises safeguard the machine learning models behind their most important products. HiddenLayer is the only company to offer turnkey security for AI that does not add unnecessary complexity to models and does not require access to raw data and algorithms. Founded by a team with deep roots in security and ML, HiddenLayer aims to protect enterprise's AI solutions from inference, bypass, extraction attacks, and model theft. The company is backed by a group of strategic investors, including M12, Microsoft's Venture Fund, Moore Strategic Ventures, Booz Allen Ventures, IBM Ventures, and Capital One Ventures.

LEARN MORE:

www.hiddenlayer.com

FOLLOW US:

Research Twitter LinkedIn

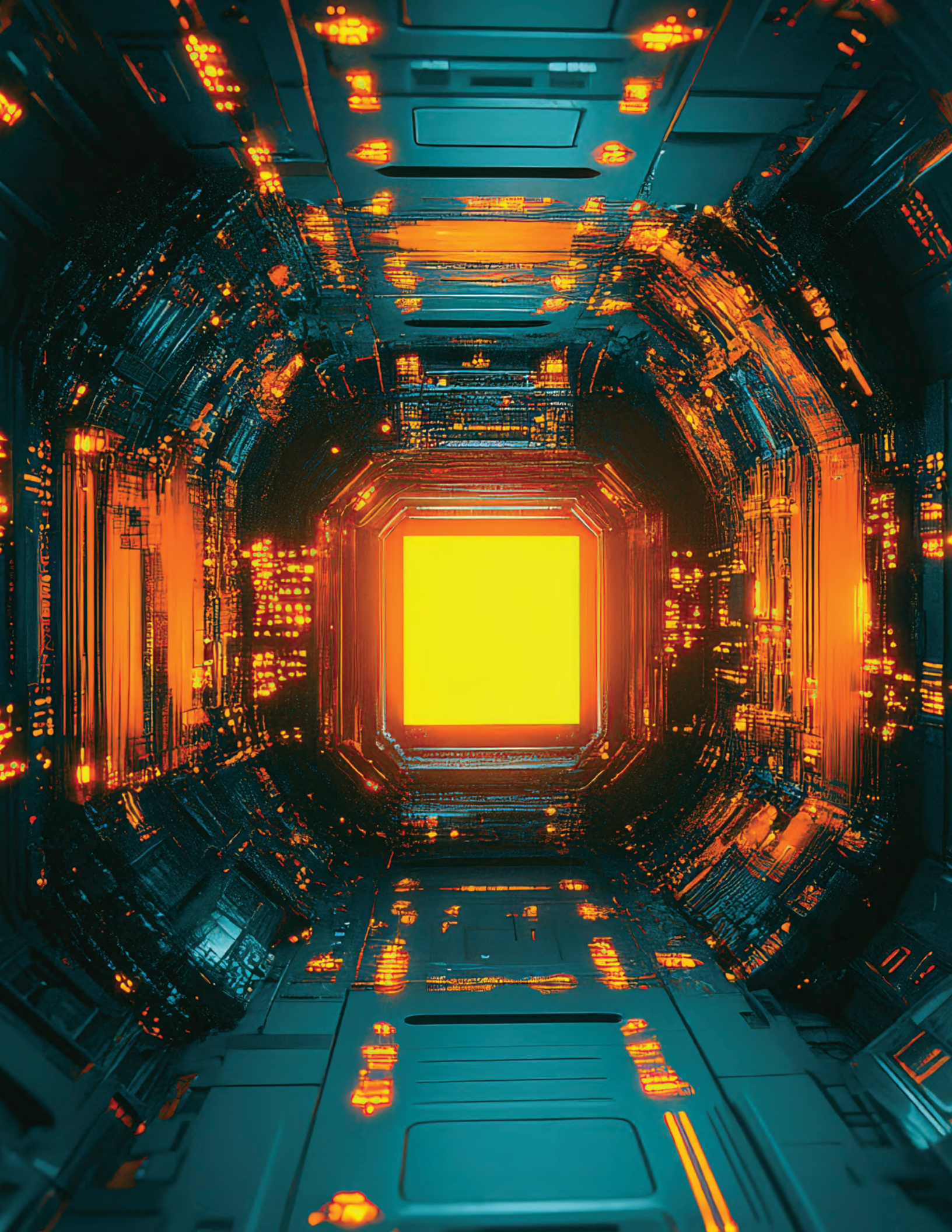
REQUEST A DEMO:

https://hiddenlayer.com/book-a-demo/

AUTHORS/CONTRIBUTORS

A special thank you to the teams that made this report come to life:

Marta Janus, *Principal Security Researcher*
Eoin Wickens, *Technical Research Director*
Tom Bonner, *SVP, Research*
Malcolm Harkins, *Chief Security & Trust Officer*
Jason Martin, *Director, Adversarial Research*
Travis Smith, *VP of ML Threat Operations*
Ryan Tracey, *Principal Security Researcher*
Jim Simpson, *Threat Operations Specialist*
Samantha Pearcy, *Manager of Content Strategy*
Kristen Tarlecki, *VP of Marketing*
Arman Abdulhayoglu, *Director of Product Marketing*
Kieran Evans, *Principal Security Researcher*
Kevin Finnigin, *Principal Security Researcher*
Marcus Kan, *AI Security Researcher*
Ravi Balakrishnan, *Principal Security Researcher*
Kenneth Yeung, *AI Threat Researcher*
Kasimir Schulz, *Director, Security Research*
Megan David, *AI Researcher*




HIDDENLAYER