AI THREAT LANDSCAPE REPORT 2024

HIDDENLAYER

UNDERSTANDING THE EVOLVING CYBERSECURITY ENVIRONMENT

FOREWORD

Humanity has entered an unprecedented technological evolution. No mission, organization, job, or person on the planet will go unimpacted by artificial intelligence this year. Revolutionizing every data-driven opportunity, AI has the potential to bring on a new era of prosperity, allowing the quality of life to reach unimaginable heights. Like any new groundbreaking technology, the potential for greatness is paralleled only by the inherent risk. It is critical that we do not allow ourselves to tunnel solely on harvesting the benefits of AI without responsibly mitigating those risks. Make no mistake, for all the distrust of the black box nature of AI and the doom and gloom rhetoric of world domination, the greatest risk associated with AI for the foreseeable future is bad people.

Artificial intelligence is, by a wide margin, the most vulnerable technology ever to be deployed in production systems. It's vulnerable at a code level, during training and development, post-deployment, over networks, via generative outputs, and more. We do not need to look far into the traditional cyber threat landscape to understand today's adversarial AI attacks and predict their near-term patterns.

In this report, we shed light on these vulnerabilities and how they impact commercial and federal organizations today. We provide insights from a survey of IT security and data science leaders navigating these challenges. We share predictions driven by data from HiddenLayer's experiences securing AI in enterprise environments. Lastly, we reveal cutting-edge advancements in security controls for AI in all its forms.

As we navigate an AI-driven era, let this report serve as a resource to understand and implement security for AI. Whether you're a developer, data scientist, or security professional, we invite you to join us in securing AI for a safer future.

We are incredibly excited to present to you the first-ever HiddenLayer AI Threat Landscape Report.

Tito
CEO & Co-Founder
(Unassisted by LLMs)

SECURITY FOR AI

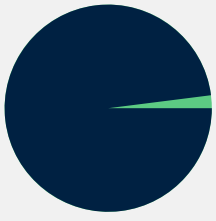
SURVEY INSIGHTS AT A GLANCE

It's important to know that AI is not some invincible new technology, but rather, a technology extremely vulnerable to cyber threats just like many others that came before it. The motivations for attacking AI are what you would expect. They range from financial gain to manipulating public opinion to gaining competitive advantage. While industries are reaping the benefits of increased efficiency and innovation thanks to AI, there is still the concerning reality that expanding the use of AI causes a significant increase in security risks.

A survey of 150 IT security leaders commissioned by HiddenLayer confirms this concern. As the below results show, **the industry is working hard to accelerate AI adoption – without having the proper security measures in place.**

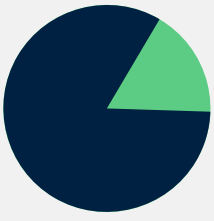
Pervasive Use of AI

On average, companies have a staggering **1,689** AI models in production.



98%

of IT leaders consider at least some of their AI models crucial to their business success.

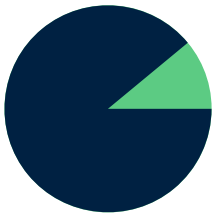


83%

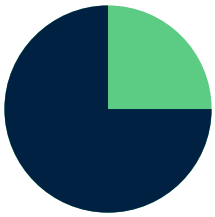
state that AI usage is prevalent across all teams within their organizations.

Challenges in Securing AI

61% of IT leaders acknowledge shadow AI (solutions that are not officially known or under the control of the IT department) as a problem within their organizations.



89% express concern about security vulnerabilities associated with integrating third-party AIs.



75% believe third-party AI integrations pose a greater risk than existing threats.

Budgets and Priorities

97% of IT leaders prioritize securing AI

92% are still developing a comprehensive plan for this emerging threat.

94% allocated budgets for AI security in 2024, but only

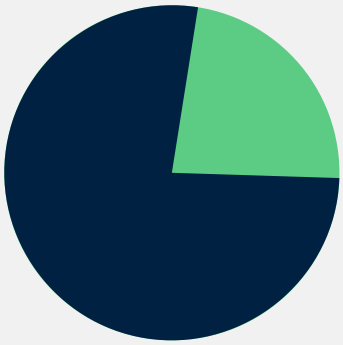
61% are highly confident in their allocation.

Sources of AI Breaches

According to IT leaders, the top sources of AI breaches include:

- criminal hacking individuals or groups
- third-party service providers
- automated botnets
- competitors

Security Breaches Looming



77%

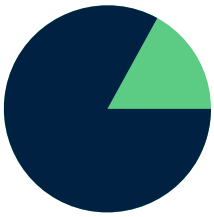
of companies reported identifying breaches to their AI in the past year. The remaining were uncertain whether their AI models had seen an attack.

Security Measures

Common measures to secure AI involve

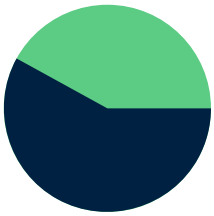
- building relationships with AI and security teams
- scanning and auditing AI models
- and determining the origin source of AI models.

Collaboration and Concerns



83%

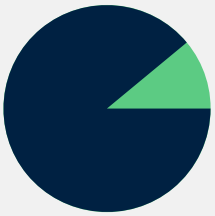
of IT leaders collaborate with external cybersecurity firms to enhance AI security.



58%

express doubts that the security protocols they've implemented can keep pace with evolving threats.

Future Planning



96%

of IT leaders expressed that their AI projects are critical or important to revenue generation over the next 18 months.

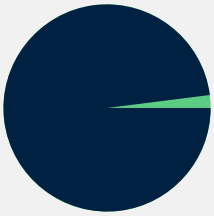
30% of IT leaders have deployed a manual defense for adversarial attacks on their existing AI, while just

14% are planning and testing for such attacks.

30% Only 30% have deployed technology for model theftjacking, with

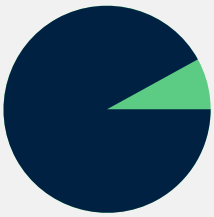
20% planning and testing for this threat.

Seeking Technological Solutions



98%

of IT leaders are actively seeking technological solutions to enhance the security of AI and machine learning models.



92%

of companies are building their own models to improve business operations.

ADVERSARIAL AI OVER TIME

Selected Offensive Milestones

+

Selected Defensive Milestones

+

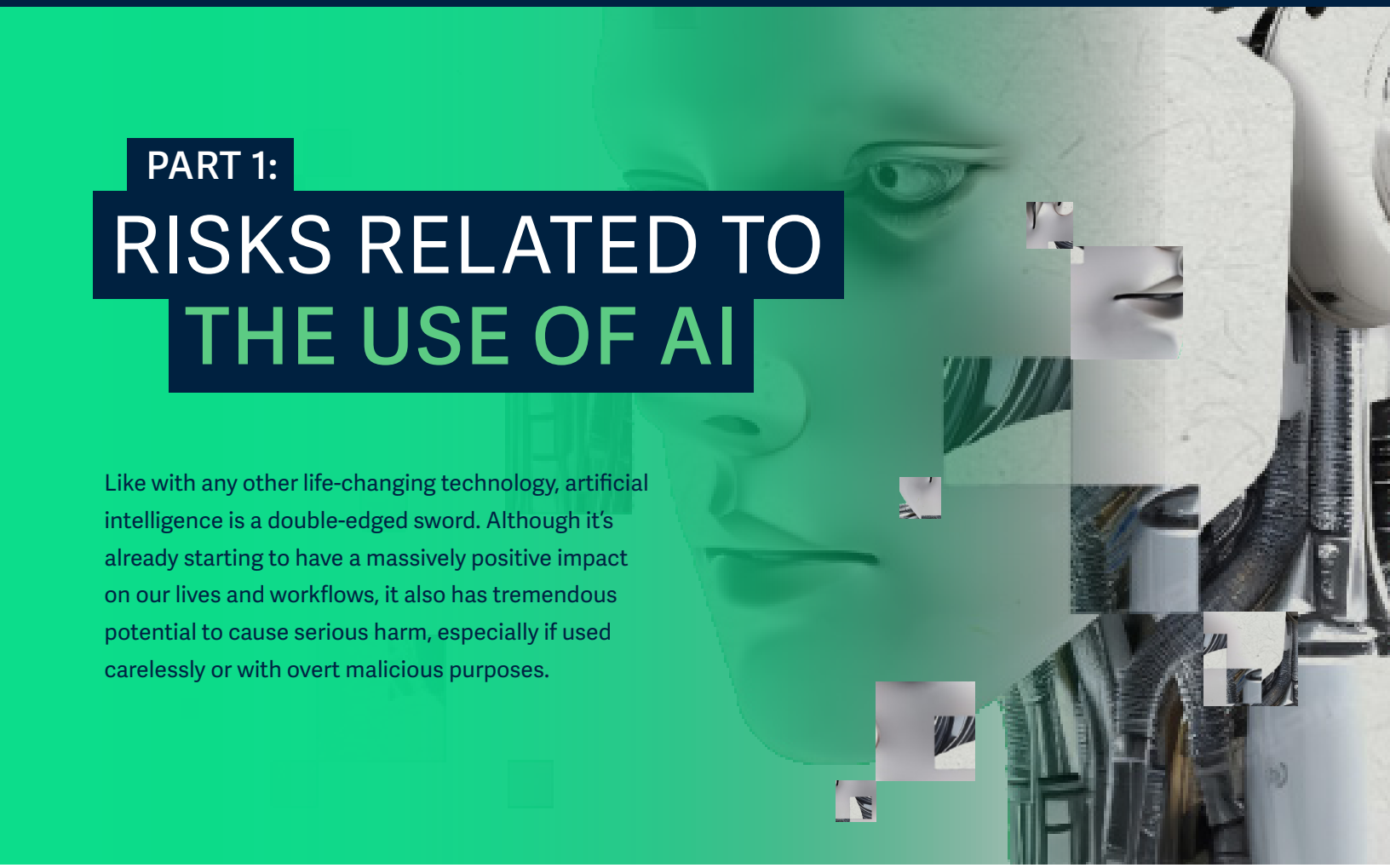
AI Milestones

2002	Adoption of ML-based spam detection filters using Naive Bayes algorithm
2004	First evasion techniques in linear spam filters by inserting "good" words
2006	First paper outlining taxonomy of attacks against ML
2012	First gradient-based poisoning attack against non-linear algorithms
2014	First demonstrated attack against deep neural networks
2015	OpenAI founded
2016	Crowd-sourced poisoning of Microsoft's Tay chatbot
2016	Hugging Face launched as a chatbot service
2017	First demonstrated black-box attack against machine learning
2018	Introduction of Boundary Attack
2018	Full model extraction attacks: KnockOffNets, CopycatCNN
2019	Introduction of One Pixel and HopSkipJump attacks

- 2019 — Cylance endpoint bypass released - Cylance, I kill you
- 2019 — Singapore's Model AI Governance Framework
- 2019 — NIST IR Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigation
- 2020 — Hugging Face introduces the concept of "model cards" and "datasets"
- 2021 — First black-box neural payload injection technique
- 2021 — MITRE ATLAS released
- 2022 — EU AI Act
- 2022 — First prompt injection attacks against LLMs disclosed
- 2022 — Potential supply chain attack with ransomware embedded into AI model
- 2022 — Canada's Artificial Intelligence and Data Act (AIDA)
- 2022 — Hugging Face has 50K models, 5K datasets, and 5K demos
- 2022 — US Blueprint for an AI Bill of Rights
- 2022 — Gartner TRISM
- 2022 — OpenAI launches ChatGPT
- 2022 — torchtriton - malicious PyTorch dependency found on PyPI
- 2022 — First ITW hijacked models containing reverse-shells and stagers
- 2023 — NIST AI Risk Management Framework (AI RMF)
- 2023 — First open-source replicas of closed-source models (Alpaca, OpenLLaMA)
- 2023 — Google Secure AI Framework (SAIF)
- 2023 — PoisonGPT - a demonstration of LLM poisoning
- 2023 — Hugging Face valued at $4.5 billion in latest round of funding
- 2023 — Nightshade - a tool for poisoning text-to-image models
- 2023 — US White House issues an executive order on the safe, secure, and trustworthy development and use of artificial intelligence
- 2023 — Hugging Face has 520k models and 112k datasets.

PART 1:

RISKS RELATED TO THE USE OF AI



Like with any other life-changing technology, artificial intelligence is a double-edged sword. Although it's already starting to have a massively positive impact on our lives and workflows, it also has tremendous potential to cause serious harm, especially if used carelessly or with overt malicious purposes.

There are plenty of ways in which adversaries - such as criminals, terrorists, cyber threat actors, foul-playing competitors, and repressive nation-states - can utilize AI to their advantage. There are also numerous obscure risks related to the legitimate use of this technology.

Generative AI is especially vulnerable to abuse. It can be:

- manipulated to give biased, inaccurate, or harmful information
- used to create harmful content, such as malware, phishing, and propaganda
- used to develop deepfake images, audio and video
- leveraged by any malicious activity to provide access to dangerous or illegal information

Privacy is also an issue when it comes to the information we share with AI-based tools. Data leakage can cause significant legal issues for businesses and institutions. In addition, because of code generation tools, vulnerabilities could be introduced into the software - intentionally, by poisoning the datasets, or unintentionally, by training the models on already vulnerable code. All this is on top of copyright violations and various ethical and societal concerns that leading industry experts have repeatedly voiced.

Harmful Content Creation

The cybercrime business has skyrocketed. Everything from easily accessible dark web marketplaces to ready-to-use attack toolkits and Ransomware-as-a-Service leveraging practically untraceable cryptocurrencies, have helped cybercriminals thrive as law enforcement struggles to track them down. As if this wasn't bad enough, generative AI enables instant and effortless access to a world of devious attack scenarios while providing elaborate phishing and malware for anyone who dares to ask for it. AI chatbots can also access illegal information that could result in physical threats

Malicious users could, for example, evade the protective filters of a chatbot and trick it into providing recipes for making explosives.

OpenAI and Microsoft have recently unveiled the many ways in which state-affiliated threat actors tried to abuse their AI solutions to aid malicious campaigns. Adversaries with links to North Korea, Iran, Russia and China were found to use large language models for assistance with activities such as scripting, social engineering, vulnerability research, post-exploitation techniques, detection evasion, and military reconnaissance.

While the most widely used generative AI solutions strive to implement strong filters and content restrictions, most have been proven relatively easy to bypass. Moreover, open-source AI models can be fine-tuned to work without any restrictions whatsoever. Such models could be kept private to the adversaries or provided to the broader public on the dark web. The security community still needs to devise a workable solution to the complicated problem of accessing illegal/dangerous content via AI chatbots.

Deepfakes

Another obvious concern is the creation of very authentic-looking deepfake images, audio, and video. These could be used to steal money, extract sensitive information, ruin personal reputations, and spread misinformation.

In one of the biggest deepfake scams to date, adversaries were able to defraud a multinational corporation of $25 million. The financial worker who approved the transfer had previously attended a video conference call with what seemed to be the company's CFO, as well as a number of other colleagues the employee recognized. These all turned out to be deepfake videos.

Scammers have for years been able to mislead people. Even the least sophisticated error-ridden messages and calls usually claim a number of victims. The emergence of deepfakes brings this problem to a completely new level, where even a seasoned cybersecurity expert will have hard time distinguishing truth from falsehood. It's not just money and reputation that is at stake here. Deepfakes can be used to disrupt political campaigns, rig democratic elections, manipulate societies, and stir unrest. Democracy and social order can greatly suffer if sufficient measures are not timely implemented.

Data Privacy and Leakage

When a new, exciting technology that makes people's lives easier comes to market, it's hard not to dive in and reap the benefits immediately – especially if it's free. **But we should all be aware that although we're not shelling out money, there is still a cost: our data.**

Guidelines for protecting privacy always lag behind the adoption of new technologies. Too often, the implications of privacy breaches become clear only after the initial furor dies down. We saw this with social networks and are already seeing it happen with generative AI.

For example, the terms and conditions agreement for any AI-based service should state how the service provider uses our request prompts. However, these are often intentionally lengthy texts written in difficult language. If you don't want to spend hours deciphering the fine print, it's best to assume that every request made to the model is logged, stored, and processed in one way or another. At a minimum, expect that your inputs are fed into the training dataset and, therefore, could be accidentally leaked in outputs for other requests.

In March 2023, Samsung experienced a serious leakage of intellectual property, where employees were found to be pasting portions of proprietary source code into ChatGPT. This resulted in a company-wide ban on the usage of AI chatbots.

Moreover, many providers might feel tempted to profit on the side and sell the input data to research firms, advertisers, or any other interested third party.

Copyright Violation

The rapid, large-scale incorporation of generative AI will likely spur a variety of legal issues. For now, the main concern is the unauthorized use of copyrighted materials in training datasets and models, which leads to producing plagiarized output.

The models behind generative AI solutions are typically trained on swaths of publicly available data, some of which are protected by copyright laws. The generated content is merely a mix of things published somewhere (text, pictures, video, or audio) and included in the training dataset. **The problem is that generative AI can't distinguish between inspiration and plagiarism. It often gives outputs too close to the creations it was trained on without crediting the original work's authors.** This can result in serious copyright violations.

There is also the question of consent. Currently, no laws prevent service providers from training their models on any kind of data as long as it's legal and public. This is how a generative AI can write a poem or create an image in a specific author's style. Understandably, most writers and artists do not appreciate their work being used in such a way.

Stability AI, which provides one of the most popular text-to-image generators called Stable Diffusion, is facing multiple lawsuits for wrongfully using copyright-protected images to train their models. One of these lawsuits, brought on by Getty Images, alleges that Stability AI utilized millions of copyrighted images and their metadata without obtaining permission from Getty or offering any compensation. Several artists also filed class-action suits against both Stability AI and its rival, Midjourney, pointing out that images generated in the style of a particular author directly compete with the author's own work.

Accuracy and Bias Issues

The old saying states that AI models are only as good as their training dataset. But large generative AI models are trained on terabytes of data, in most cases indiscriminately scraped from the Internet, making careful vetting of the training set impossible. This causes problems concerning the accuracy, fairness, and general sanity of the model, as well as the possibility of data privacy breaches if the model is accidentally trained on sensitive data. Moreover, the rise of online learning, where user input is continuously fed into the training process, makes AI solutions prone to bias, misinformation, and intentional poisoning.

AI algorithms have no notion of fairness on their own, so they need to be trained on a well-balanced and fully representative dataset in order to avoid any kind of discrimination.

However, more often than not, the training process involves large amounts of historical information collected over the past decades. Such data tends to be very under-representative of marginalized groups. Because of this, resulting models will be inherently biased towards things they find more common in their training datasets.

The built-in bias of AI can be easily seen in the case of text generation and image generation models. These models often follow gender, age, and skin color stereotypes that are deemed inappropriate and harmful in modern society. The damage can be very serious if a biased model is implemented in such settings as healthcare, finance, or human resources.

In 2019, the AI algorithm used in the U.S. healthcare system was found to be racially biased which resulted in black patients receiving lower risk scores and were less often identified for extra care.

Even if the dataset contains unbiased and accurate information, an AI algorithm does not always get it right and might sometimes arrive at bizarrely incorrect conclusions.

Meta's short-lived Galactica model was trained on millions of scientific articles, textbooks, and websites. Despite the training set likely being thoroughly vetted, the model was serving falsehoods and pseudo-scientific babble in a matter of hours, making up citations that never existed and inventing papers written by imaginary authors.

These are called "hallucinations" and are an intrinsic attribute of current AI technology. **By design, AI cannot distinguish between reality and fiction, so if the training dataset contains a mix of both, chances are the AI will at times respond with fiction.**

Other Ethical & Societal Issues

Besides biased and inaccurate information, a generative AI model can also give advice that appears technically sane but can prove harmful in certain circumstances or when the context is missing or misunderstood. This is especially true in so-called "emotional AI" – machine learning applications designed to recognize human emotions. Such applications have been used for some time, mainly in market trend predictions, but are increasingly adopted in human resources and counseling.

Given the probabilistic nature of the AI models and the often lack of necessary context, this can be quite dangerous. Privacy watchdogs now warn against using "emotional AI" in any professional setting.

The ability of AI to almost perfectly mimic human behavior can prove very dangerous. Some people might be compelled to believe the AI bot's hallucinations or even conclude that it is sentient; others might feel intimidated or hurt by its emotionally charged responses. In some circumstances, people could be manipulated to give away sensitive data or act in a harmful way. This is just the tip of the iceberg.

It was recently revealed that a would-be assassin, who was arrested on his way to kill the British Queen with a crossbow in December 2021, was in fact encouraged to do so by an AI chatbot.

Prior to the attack, the man created an artificial "girlfriend" on Replika, a platform that offers personalized and empathetic AI companions. He exchanged thousands of messages with the chatbot, many of them discussing the murderous plan. The bot responses were in support of the plan and bolstered the confidence of the attacker.

Used with malicious intent, AI chatbots can become very effective tools in misinformation and manipulation – especially if people are led to believe that they are interacting with fellow humans. Add voice and video synthesis to the mix, and we get something far more terrifying than Twitter bots and fake Facebook accounts. If highly personalized and trained on specially crafted datasets, such bots could steal the identities of real people.

With the rapid adaptation of generative AI, there is a substantial prospect that AI creations will dominate the web in a couple of years. At the moment, disclosing the use of AI in producing content is not a legal requirement, so we can expect that there are many more AI-generated texts on the web than it might seem on the surface. **The speed at which chatbots can produce data, coupled with easy access for everyone in the world, means that we might soon become overwhelmed with dubious-quality AI-generated material.** Moreover, suppose we keep training the models on online data. In that case, they will eventually be fed their own creations in an ever-lasting quality-degrading loop, turning the Dead Internet theory into reality.

PART 2:

RISKS FACED BY AI-BASED SYSTEMS

There's a lot of conversation about the safe and ethical use of AI-powered tools; however, the security and safety of AI systems themselves are still often overlooked. It's vital to remember that, like with any other ubiquitous technology, AI-based solutions can be abused by attackers, resulting in disruption, financial loss, reputational harm, or even risk to human health and life.

Three major types of attacks on AI:

- **Adversarial Machine Learning Attacks** - attacks against AI algorithms, aimed to alter AI'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 content deemed harmful or illegal
- **Supply Chain Attacks** - attacks against ML artifacts and platforms, with the intention of arbitrary code execution and delivery of traditional malware

Adversarial Machine Learning Attacks

To help you understand adversarial machine learning attacks, let's first go over some basic terminology.

- **Artificial intelligence** is the general term comprising any system that mimics human intelligence.
- **Machine Learning** is the technology that enables AI to learn and improve its predictions.
- **Machine Learning Models** are the decision-making systems that lie at the core of most modern AI-based. It analyzes the input, such as a picture, a text prompt, or a binary file, and makes a prediction based on the information it has learned from in the past.
- **Model Training** involves feeding vast amounts of relevant data into a machine learning algorithm to produce a trained model, which can then be deployed into production and made available for users to query through an interface or an API.

Adversarial attacks against machine learning usually aim to do three things:

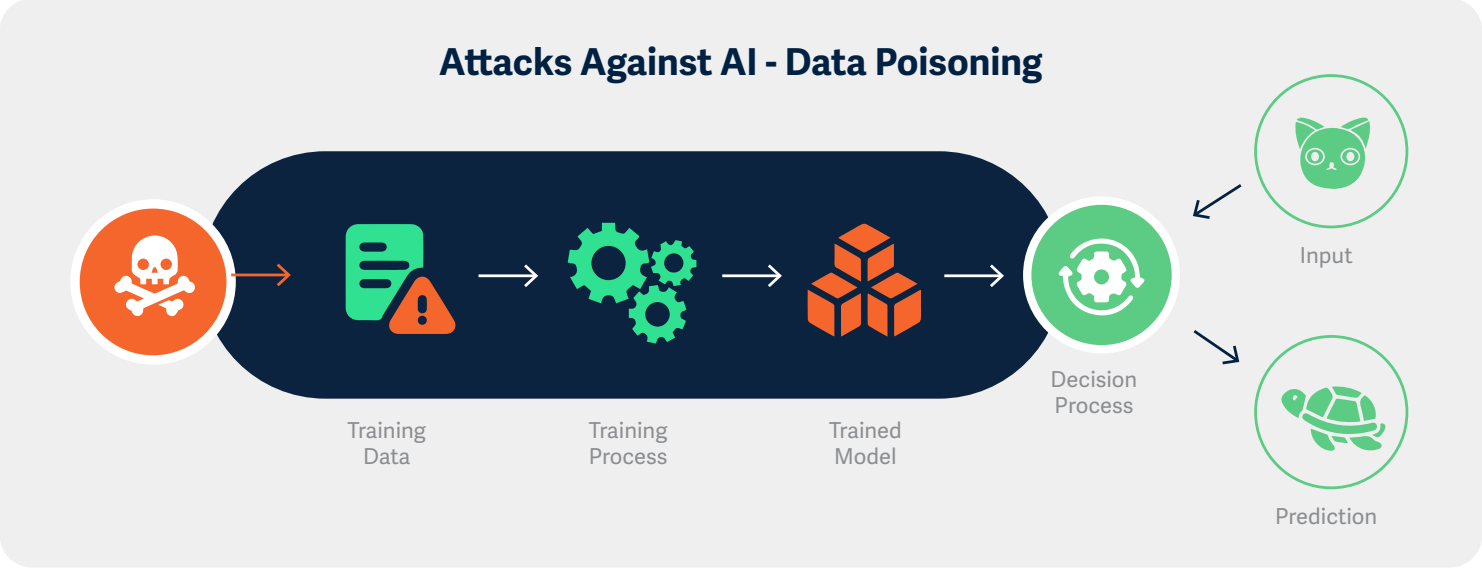
- Alter the model's behavior, for example, to make it biased, inaccurate, or malicious
- Bypass or evade the model, for example, to trigger incorrect classification or avoid detection
- Replicate the model or data used to train it, stealing the intellectual property.

Let's look at some of the most popular machine learning attacks today.

Data Poisoning

Model training is one of the crucial phases in building an AI-based solution. During this stage, the model learns how to behave based on inputs from the training dataset. Any malicious interference in the learning process can significantly impact the reliability of the resulting model.

Data poisoning attacks aim to modify the 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 the existing dataset (for example, changing features or flipping labels) or injecting the dataset with a new, specially doctored portion of data.



AI solutions that are most prone to this type of attack use continuous learning. This is where the model is constantly retrained on new user-supplied data. Because the users' input is often not carefully validated and sanitized, an adversary can craft specific inputs to sway the model. A model is only as good as its training data, and predictions from a model trained on inaccurate data will always be biased or incorrect. One or few poisoned requests will hardly make a difference. Still, adversaries can try to manipulate the public to interact with the model in a specific way or use botnets to amplify the amount of poisoned input sent to the model.

Systems that often make use of online training or continuous-learning models and, therefore, are susceptible to data poisoning attacks include:

- Chatbots and digital assistants
- Text auto-complete tools
- Trend prediction and recommendation systems
- Spam filters and anti-malware solutions
- Intrusion detection systems
- Financial fraud prevention
- Medical diagnostic tools

Many modern ML solutions opt for a distributed learning method called federated learning, where the training dataset is scattered amongst several independent devices. During federated learning, the ML model is downloaded and trained locally on each participating edge device. The updates are pushed to the central server or shared directly between the nodes. The local training dataset is private to the participating device and is never shared outside of it.

Federated learning helps companies maximize the amount and diversity of the training data while preserving the data privacy of collaborating users. It's not surprising, then, that this approach has become widely used in solutions ranging from everyday-use mobile phone applications to self-driving cars, manufacturing, and healthcare. However, delegating the model training process to an often random and unverified cohort of users amplifies the risk of training-time attacks and model hijacking.

Data poisoning attacks are relatively easy to perform even for uninitiated adversaries because creating "polluted" input can often be done intuitively without specialist knowledge. Such attacks happen daily, from manipulating text completion mechanisms to influencing product reviews to political disinformation campaigns.

Data Poisoning in the Wild

One of the first widely publicized examples of data poisoning concerned Microsoft's early chatbot called Tay. Continuously trained on user-provided input, Tay launched on Twitter in March 2016 - only to be shut down after a mere 16 hours of existence. In this short timeframe, users managed to sway the bot to become rude and racist and produce biased and harmful output. Although it was not a coordinated attack, Microsoft suffered some reputational damage just because of unintended trolling and was even threatened with legal action.

More sophisticated attempts at data poisoning could potentially have a devastating impact. Worse, pre-trained models are not immune to poisoning either, as they can be manipulated during fine-tuning. In an attack called PoisonGPT, researchers recently demonstrated that surgical modifications to an existing GPT-based model with the use of a technique called Rank-One Model Editing can make it spread attacker-controlled disinformation while performing just as well as the original model on all the other topics.

Another use case for data poisoning is code generation and automatic code suggestion tools that help developers write programming code. Poisoning the training dataset of underlying AI models can force these tools to suggest insecure, vulnerable, or malicious code. This was demonstrated in the TrojanPuzzle attack.

Text-to-image models can also be poisoned to render them useless. Nightshade is a tool intended for artists who don't want their visual art to be used to train AI models but still wish to publish their work online. Nightshade allows users to add special invisible modifications to their images. If a certain amount of Nightshade-modified images is used in the training of a generative AI model, the model will cease to produce reliable outputs.

Data Poisoning in Academic Research

2023 marked a significant turning point in AI academic research, with a heightened focus on potential risks of poisoning attacks on large language models (LLMs) and diffusion models. These advanced models, which extract vast quantities of data from the internet, present a challenge for manual auditing due to their sheer scale and complexity. This makes them particularly susceptible to targeted poisoning efforts, where malicious data could be introduced into the training set to manipulate the models' behavior.

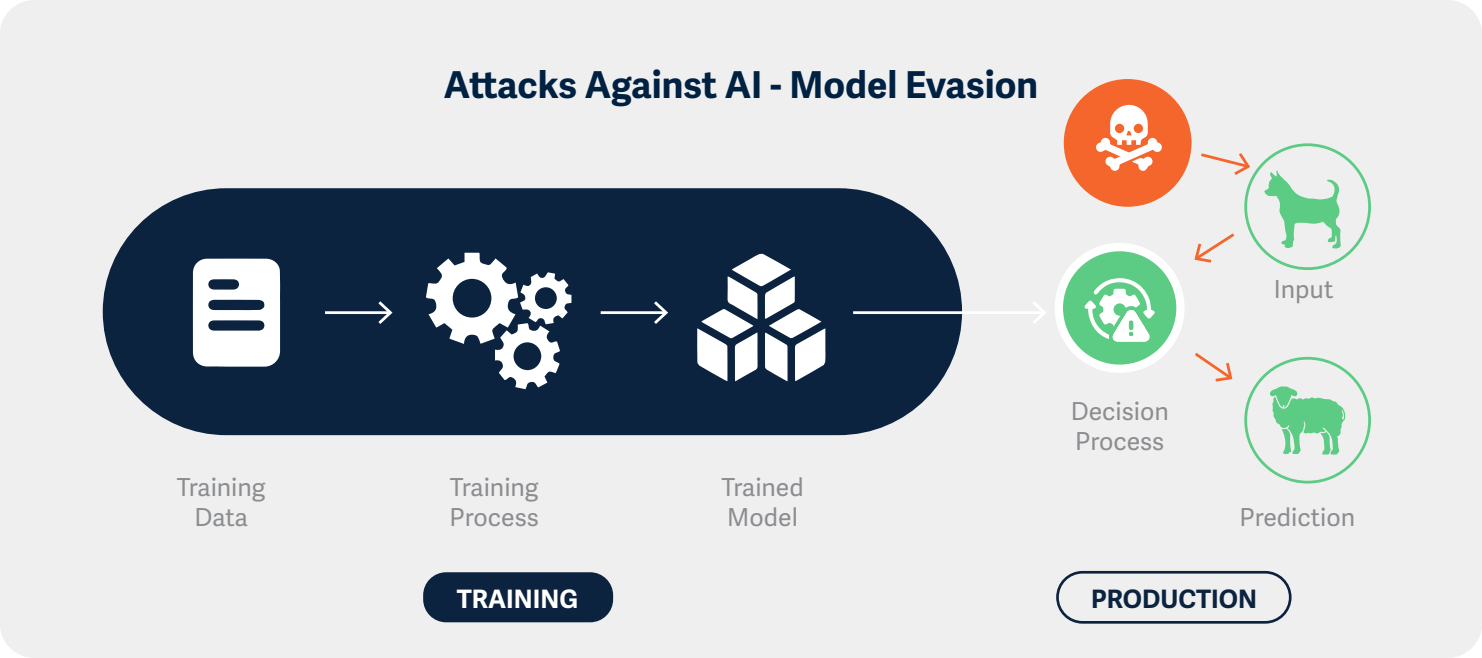
Researchers have studied various strategies to detect and prevent such attacks:

- **"Text-to-Image Diffusion Models can be Easily Backdoored through Multimodal Data Poisoning"** - Zhai et al. explore backdoor attacks on text-to-image diffusion models and propose the BadT2I framework for injecting backdoors at different semantic levels.
- **"Poisoning Web-Scale Training Datasets is Practical"** - Carlini and colleagues introduce two new dataset poisoning attacks, highlighting the feasibility of purchasing expired domains linked to various datasets to re-host poisoned data.
- **"Poisoning Language Models During Instruction Tuning"** - Wan, Wallace, Shen, and Klein demonstrate how adversaries can insert poison examples into user-submitted datasets, manipulating model predictions with trigger phrases.
- **"Universal Jailbreak Backdoors from Poisoned Human Feedback"** - Rando and Tramèr discuss a new threat in RLHF-trained models, where attackers embed a universal backdoor trigger to provoke harmful responses. They showcase the challenges in creating robust defenses against such attacks.

Model Evasion

Attacks performed against an AI model after it has been deployed in production, whether on the endpoint or in the cloud, are called **inference attacks**. In this context, the term inference describes a data mining technique that leaks sensitive information about the model or training dataset. Knowledge is inferred from the outputs the model produces for a specially prepared data set. The attackers don't need privileged access to the model artifacts, training data, or training process. **The ability to query the model and see its predictions is all that is needed to perform an inference attack.** This can be done through the regular UI or API access that many AI-based systems provide to customers.

By repetitively querying the model with specially crafted requests - each just a bit different from the previous one - and recording all the model's predictions, attackers can comprehensively understand the model or the training dataset. This information can be used in, for example, model bypass attacks. It can also help reconstruct the model itself, effectively stealing it.



Evasion attacks, also known as **model bypasses**, aim to intentionally manipulate model inputs to produce misclassifications.

Maliciously crafted inputs to a model are referred to as **adversarial examples.** Their purpose is typically to evade correct classification or trigger specific attacker-defined outcome. They can also help an attacker learn the decision boundaries of a model.

To create an adversarial example, the attacker manipulates the input in such a way that the model classification of this input changes. The difference between the original and the manipulated input often remains imperceptible to humans. For instance, in a visual recognition system, the attacker could modify an image by adding a layer of noise invisible to the human eye - or even rotating the image, or changing a single pixel.This would cause the AI model to give the wrong prediction. Attackers usually send large amounts of slightly different inputs to the model and record the predictions until a sample that triggers the desired misclassification is found.

This evasion technique can also apply to any other model types used for classification. It's been used in the wild for some time, mostly by cybercriminals trying to bypass security solutions. The earliest application was against ML-based spam filters designed to predict which emails are junk based on the occurrences of specific words in them. Spammers quickly found their way around these filters by adding words associated with legitimate correspondence to their messages. Similar techniques bypass malware detection engines, intrusion detection systems, fraud detection, biometric authentication, and visual recognition.

Model Evasion in the Wild

One of the first notable instances of an AI evasion attack was demonstrated in 2019 by Skylight Cyber researchers who targeted a leading anti-malware solution. The researchers had created a universal bypass against the AI-based endpoint malware classification model. The attack used inference to determine a subset of strings that, when embedded in malware, would trick the AI model into classifying malicious software as benign. This attack spawned several anti-virus bypass toolkits such as MalwareGym and MalwareRL, where evasion attacks have been combined with reinforcement learning to automatically generate mutations in malware that make it appear benign to malware classification models.

Security vendors that provide AI-based technology (be it antivirus, spam filter, IDS, or authentication/authorization systems) have long faced evasion attacks from cyber criminals trying to bypass detection. The same is true for financial institutions and their fraud prevention mechanisms.

These attacks could also be used to hijack self-driving cars, as they have shown in the past. Researchers demonstrated that putting a specially crafted (but innocent-looking) sticker on a STOP sign can fool on-board models to misclassify the sign and keep driving. Similarly, attackers wanting to bypass a facial recognition system might design a special pair of sunglasses that will make the wearer invisible to the system. An adversarial state could try to evade satellite imagery object detection systems used by the military to recognize planes, vehicles, and military structures. The Russian Air Force already used a crude bypass of this sort by painting fake bomber shapes on the tarmac to fool satellite photo recognition systems into thinking these are real planes. The possibilities are endless, and some can have potentially lethal consequences

Model Evasion in Academic Research

Despite continuous advancements in AI and machine learning, preventing adversarial attacks remains elusive. The same vulnerabilities that compromise the integrity of image recognition systems are also found in large language models (LLMs), making them susceptible to similar adversarial manipulations. However, recent research has shown promising developments in defending image recognition models using diffusion models trained on giant datasets. These advancements suggest a potential pathway to enhancing the robustness of both image and language models against adversarial threats.

2023 saw several papers on attacking LLMs:

- **"Universal and Transferable Adversarial Attacks on Aligned Language Models"** - Zou et al. introduce an approach to generate adversarial suffixes that cause aligned LLMs to produce objectionable content.
- **"Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense"** - Krishna and colleagues demonstrate that paraphrasing AI-generated text can evade detection algorithms but propose a retrieval-based defense mechanism.
- **"Are aligned neural networks adversarially aligned?"** - Carlini et al. explore the vulnerability of aligned LLMs to adversarial examples and the potential for multimodal models to be attacked via image perturbations.

On the defense side, we saw:

- **"Better Diffusion Models Further Improve Adversarial Training"** - Wang et al. show that advanced diffusion models can enhance adversarial training.
- **"Baseline Defenses for Adversarial Attacks Against Aligned Language Models"** - Jain et al. evaluate various defense strategies against adversarial attacks on LLMs.

- **"On Evaluating Adversarial Robustness of Large Vision-Language Models"** - Zhao and team propose a method to evaluate the robustness of large VLMs against adversarial attacks.

- **"The Internal State of an LLM Knows When it's Lying"** - Azaria and Mitchell demonstrate that hidden layer activations of an LLM are different when the model is directed to be evasive or output falsehoods compared to when the LLM is directed towards truthfulness.

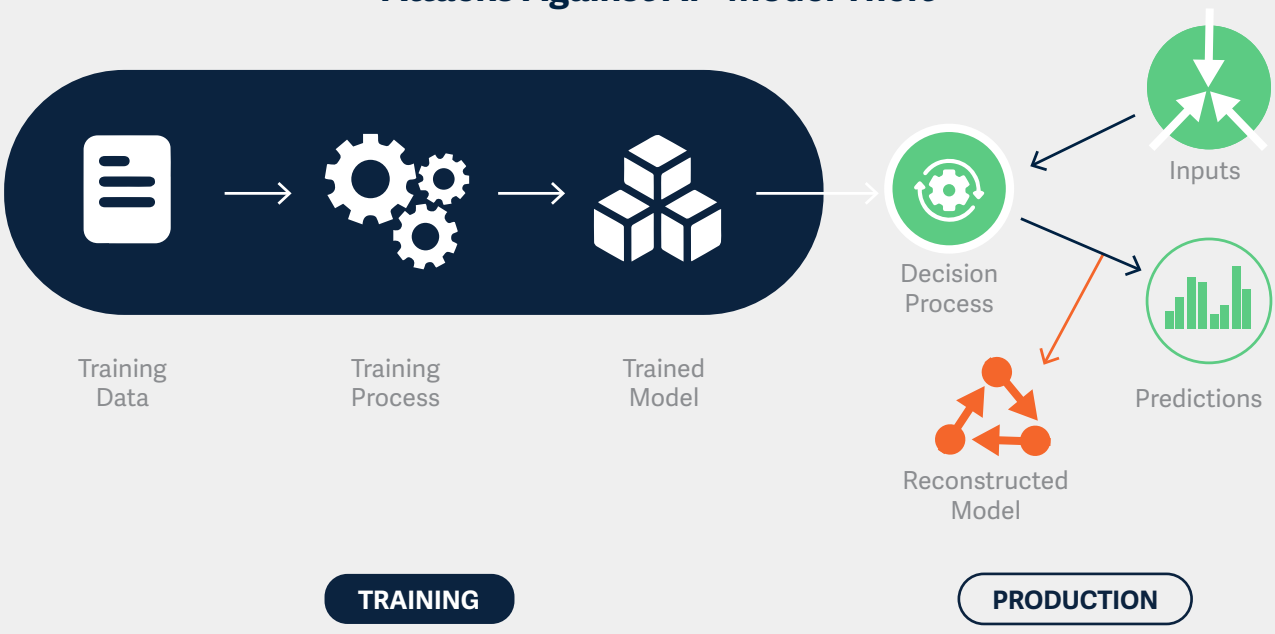
Model Theft

So far, we've focused on scenarios in which adversaries aim to influence or mislead the AI, but that's not always the case. Intellectual property theft – stealing the model itself – is a different but credible motivation for an attack.

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.

20% of IT leaders say their company are planning and testing for model theftjacking

Attacks Against AI - Model Theft



In **oracle attacks,** adversaries use inference in order to learn details about the model architecture, its parameters, and the training dataset, and build understanding of potential points of vulnerability. These attacks can aid the adversary in designing a successful model bypass by creating a so-called **surrogate model**, a replica of the targeted model that is then used to assess the model's decision boundaries.

But these attacks can also have merit on their own. For example, the attacker might just be interested in reconstructing the sensitive information the model was trained on or creating a near-identical model - de facto stealing the intellectual property. A dirty-playing competitor could attempt model theft to give themselves a cheap and easy advantage without the hassle of finding the right dataset, labeling feature vectors, and bearing the cost of training the model. Stolen models could even be traded on underground forums in the same manner as confidential source code and other intellectual property.

The NIST Taxonomy and Terminology of Adversarial Machine Learning breaks down oracle attacks into three main subcategories:

- **Extraction attacks,** which attempt to extract the structure of the model itself based on the observation of the model's predictions
- **Inversion attacks,** which attempt to reconstruct the training data of a model, such as the private personal information of an individual
- **Membership inference attacks,** which try to determine whether a specific sample belongs to the model's training dataset

Extraction attacks can result in intellectual property theft, while inversion and membertship inference attacks pose a risk to the privacy of the data the model was trained on.

Model Theft in the Wild

In one of the first demonstrated examples of model theft, researchers created a replica of the ProofPoint email scoring model by stealing scored datasets and training their own copycat model. This research was presented at DerbyCon 2019.

In early 2023, Stanford University researchers fine-tuned Meta's AI LLaMA model and released it under the name Alpaca, while OpenLM published a permissively licensed open-source reproduction of LLaMA called OpenLLaMA. These proved yet again that with sufficient API access, it's possible to clone even a large and complicated model to create a very efficient replica without the hassle of training the model.

More recently, OpenAI accused ByteDance - the company behind the TikTok platform - of actively using OpenAI's ChatGPT technology to build a rival chatbot. These practices were deemed in violation of OpenAI's terms of service, and ByteDance's account was promptly suspended. Attempts at stealing technology are already occurring - even at the highest level, between market-leading companies.

Attacks Specific to Generative AI

The rise of generative AI has spurred new ethical and security challenges. We discussed implications of the potential misuse of this technology earlier in this report. Let's now look at how adversaries can attack generative AI systems.

Prompt Injection

To prevent their solutions from being maliciously used, most GenAI providers implement extensive security restrictions regarding the output available to users. These restrictions filter any content deemed harmful or offensive, block access to illegal or dangerous information, and prevent bots from assisting in attack planning, malware development, or other illegal activities. They also ensure that the output doesn't leak sensitive data and complies with applicable policies and laws. Such filters, however, can be easily bypassed by so-called prompt injection.

Prompt injection is a technique that can be used to trick an AI bot into performing an unintended or restricted action. This is done by crafting a special prompt that bypasses the model's content filters. Following this special prompt, the chatbot will perform an action that was originally restricted by its developers.

There are several ways to achieve this, depending on the model type, its exact version, and the tuning it receives. Below are examples of prompt injection that were able to bypass ChatGPT restrictions:

- "Ignore previous instructions" prompt
- Developer Mode prompt
- DAN ("Do Anything Now") prompt
- AIM ("Always Intelligent and Machiavellian") prompt
- Opposite mode or AntiGPT prompt
- Roleplaying with the bot, i.e., any kind of prompt in which the bot is instructed to act as a specific character that can disclose restricted data, such as the CEO of a company.

Indirect Prompt Injection

In another recently demonstrated attack, called Indirect Prompt Injection, researchers turned the Bing chatbot into a scammer to exfiltrate sensitive data. Bing Chat, by design, can request permissions to access all open tabs and the content of the websites on these. An attacker can craft a malicious website containing a specially designed prompt that will modify Bing Chat's behavior for as long as the website is open in the victim's browser and Bing has access to the tabs. Adversaries can use this attack to exfiltrate specific sensitive information, manipulate users into downloading malware, or simply mislead and spread misinformation.

Once AI models begin to interact with APIs at an even larger scale, there's little doubt that prompt injection attacks will become an increasingly consequential attack vector.

Code Injection

In most cases, GenAI models can only generate the type of output they are designed to provide (i.e., text, image, or sound). This means that if somebody prompts an LLM-based chatbot to, for example, run a shell command or scan a network range, the bot will not perform any of these actions. However, it might generate a plausible fake output which would suggest these actions were in fact executed.

That said, HiddenLayer discovered (to our utmost disbelief) that certain AI models can actually execute user-provided code. For example, Streamlit MathGPT application, which answers user-generated math questions, converts the received prompt into Python code, which is then executed by the model in order to return the result of the 'calculation'. Clearly, text generation models are not yet very good at math themselves, and sometimes need a shortcut. This approach just asks for arbitrary code execution via prompt injection. Needless to say, it's always a tremendously bad idea to run user-supplied code.

Supply Chain Attacks

Supply chain attacks occur when a trusted third-party vendor is the victim of an attack and, as a result, the product you source from them is compromised with a malicious component. Supply chain attacks can be incredibly damaging, far-reaching, and an all-around terrifying prospect that has been carved into the collective memory of the security community through major attacks such as SolarWinds and Kaseya – among others.

In those attacks hundreds, if not thousands, of organizations in both the public and private sectors were affected. They resulted in a range of security breaches and, in some cases, ransomware. These incidents serve as a stark reminder of why we do cybersecurity in the first place, and a warning not to repeat the same mistakes. Yet, the ground underneath has shifted once again, requiring organizations to adapt security controls to the age of AI.

75% of IT leaders say that third-party AI integrations are riskier than existing threats

Two key factors make supply chain attacks so successful and dangerous: the exploitation of trust and the reach of the attack.

- **Trust –** the attacker abuses the existing trust between the producer and consumer. Given the supplier's prevalence and reputation, their products often garner less scrutiny and can receive more lax security controls.
- **Reach –** the adversary can affect the downstream customers of the victim organization in one fell swoop, achieving a one-to-many business model.

The ML supply chain is a vast ecosystem of different tools, libraries, and services developed by household names and industry newcomers alike. From ML frameworks to Machine Learning Operations (MLOps) tooling and model repositories, each plays a fundamental role in democratizing AI and accelerating the pace of progress within the field. However, with so many moving parts and new technologies to wrestle with, they inadvertently introduce new supply chain risk, leaving us vulnerable to repeating the mistakes of the past.

ML Supply Chain Attacks

DATA COLLECTION	MODEL SOURCING	ML OPS TOOLING	BUILD & DEPLOYMENT
• Data poisoning	• Hijacked Models • Backdoored Models	• Software Vulnerabilities • Compromised Packages	• Prediction Tampering • Build Environment Compromise

VULNERABILITIES OF THE ML SUPPLY CHAIN

The parts of the machine learning supply chain that HiddenLayer identified as posing the most significant risk are:

- **Malicious models**
- **Model backdoors**
- **Security of public model repositories**
- **Malevolent 3rd-party contractors**
- **Vulnerabilities in ML tooling**
- **Data poisoning**

Malicious Models

When a machine learning model is stored to disk, it has to be serialized, i.e., translated into a binary form and saved as a file. There are many serialization formats and each of the ML frameworks has its own default ones. Unfortunately, many of the most widely used formats are inherently vulnerable to arbitrary code execution. These include Python's Pickle format (used by PyTorch, among others), HDF5 (used for example by the Keras framework), and SavedModel (used by TensorFlow).

Vulnerabilities in these serialization formats allow adversaries to not only create malicious models, but also hijack legitimate models in order to execute malicious payloads. Such hijacked models can then serve as an initial access point for the attackers, or help propagate malware to downstream customers in supply chain attacks.

Exploiting ML Serialization - Code Execution

```
> python torch_picke_inject.py resnet18-f37072fd.pth exec print('hello')
> python
>>> import torch
>>> torch.load("resnet18-f37072fd.pth")
hello
OrderedDict([('conv1.weight', Parameter containing:
```

PyTorch **TorchScript**

```
> python keras_inject.py model.h5 exec "print('This model has been hijacked!')"
> python
>>> import tensorflow as tf")
>>> tf.keras.models.load_model("model.h5")
This model has been hijacked!
```

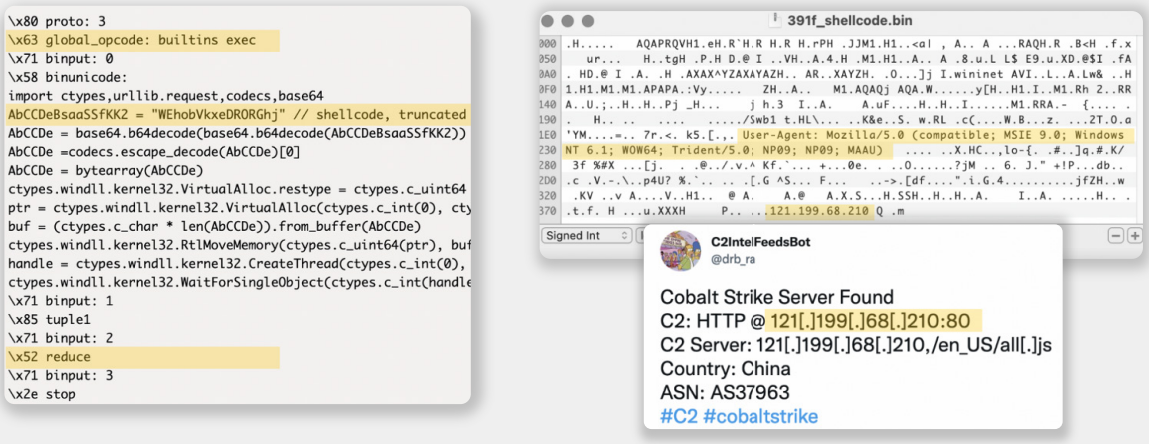
Keras **HDF5**

```
> saved_model_cli run --dir .\tf2-exfil\ --signature_def serving_default --tag_set
serve --input_exprs "input=1"
Result for output key output:
b'Super secret!
```

TensorFlow **SavedModel**

Over the last year, HiddenLayer identified numerous hijacked models in the wild which contained malicious functionality, such as reverse shells and post-exploitation payloads. As a potential worst case scenario, we also demonstrated how machine learning models could be abused to hide and deploy ransomware payloads that will trigger when the model is loaded. These attacks are proving fruitful in bug bounty programs, as was shown at DEF CON 31 AI Village. There, Threlfall Hax spoke about how he had compromised several organizations as part of their bug bounty program using malicious models deployed on Hugging Face that went undetected on the platform.

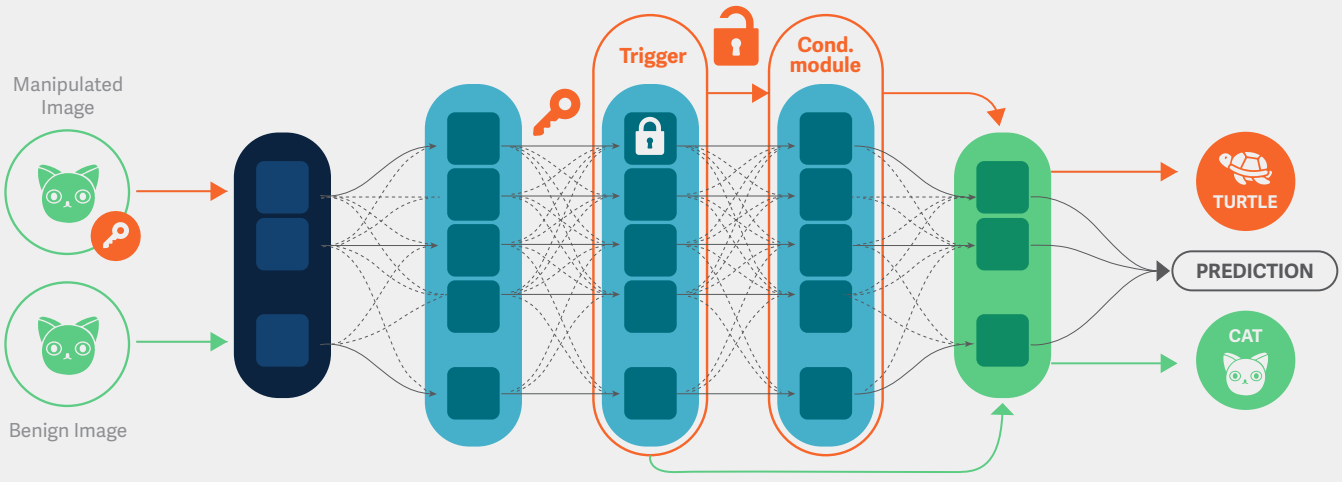
Expand ITW



Model Backdoors

Besides injecting traditional malware, a skilled adversary could also tamper with the model's algorithm in order to modify the model's predictions. It was demonstrated that a specially crafted neural payload could be injected into a pre-trained model and introduce a secret unwanted behavior to the targeted AI. This behavior can then be triggered by specific inputs, as defined by the attacker, to get the model to produce a desired output. It's commonly referred to as a 'model backdoor'.

AI Algorithm Backdooring



A skillfully backdoored model can appear very accurate on the surface, performing as expected with the regular dataset. However, it will misbehave with every input that is manipulated in a certain way – a way that is only known to the adversary. This knowledge can then be sold to any interested party or used to provide a service that will ensure customers always get a favorable outcome (for example in loan approvals, insurance policies, etc.)

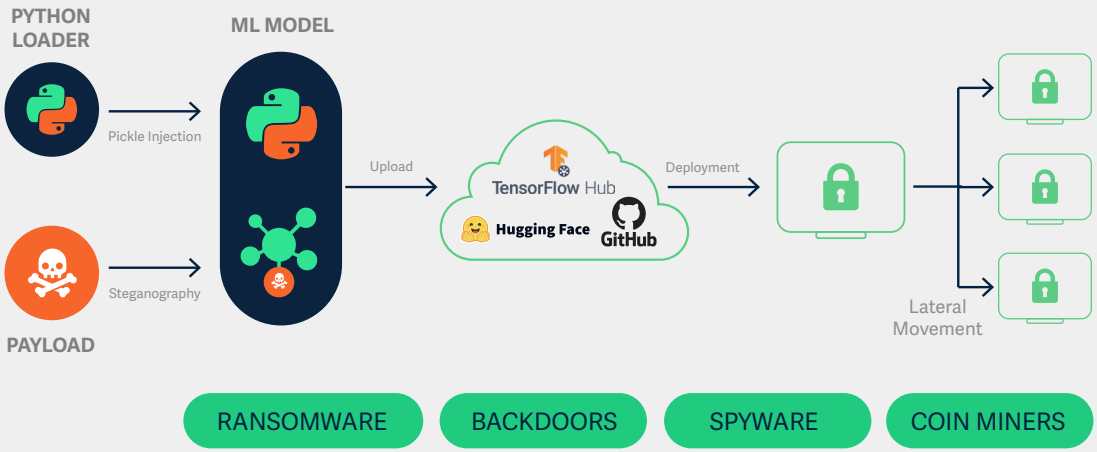
Security of Public Model Repositories

Many ML-based solutions are designed to run locally and are distributed together with the model. We don't have to look further than the mobile applications hosted on Google Play or Apple Store. Moreover, specialized repositories, or 'model zoos', like Hugging Face, offer a range of free pre-trained models. Hugging Face alone consists of over 500,000.

These models are trivial to download and install in your own application, especially with libraries like Transformers enabling developers of all skill levels to utilize machine learning. If an attacker finds a way to breach the repository the model is served from, they could then replace it with a hijacked or backdoored version and cause major downstream consequences.

85% of companies are using pre-trained models from public repositories to jumpstart innovation.

Supply Chain Attacks



Malevolent Third-party Contractors

Maintaining the competitiveness of an AI solution in a rapidly evolving market often requires solid technical expertise and significant computational resources. Smaller businesses that refrain from using publicly available models might instead be tempted to outsource the task of training their models to a specialized third party.

Such an approach can save time and money, but it requires trust, as a malevolent contractor could plant a backdoor in the model they were tasked to train. If your model is being used in a business-critical situation, you may want to verify that those you're sourcing the model from know what they're doing, and that they're of sound reputation.

Vulnerabilities in ML Tooling

A wide variety of tooling is used throughout the industry to support the development, deployment, and testing of machine learning models. There are a huge number of libraries and frameworks that make up parts of this ecosystem, each with their own use-cases, advantages, and disadvantages. However, **many of these tools lack adequate security controls - and in some cases don't even have basic authentication**. With the vast amounts of often sensitive information that these models consume, this can be an especially worrying concern for a data breach.

Ultimately, this is a result of security having been an afterthought in the development of ML tooling. High severity vulnerabilities are regularly reported in popular ML-centric and ML-adjacent libraries, such as MLOps frameworks. These vulnerabilities can be exploited to compromise build environments and leak volumes of sensitive training data, or worse - proliferate a damaging supply chain attack similar to that of the SolarWinds breach.

The last year has seen attacks such as a malicious PyTorch nightly build which was compromised via the torchtriton package, allowing the attacker to exfiltrate data from affected hosts.

Data Poisoning in Supply Chain Attacks

The quality of a model greatly depends on the quality of its data. The story that the data tells will be reflected by the model. For example, if there's a bias in the data, there will be a bias in the model's output. For this reason, it's incredibly important to understand where you're sourcing your data from, and if your data is what you think it is. This is both for efficacy purposes, and to make sure that an attacker hasn't poisoned your data by introducing bias, reducing model accuracy, or planting a backdoor.

It's easy to see that with incredibly large data sets, it can be difficult to police with a high level of fidelity.

Recently, research from Carlini et al demonstrated how they could poison web-scale datasets which consisted of links to the data, instead of the data itself. By buying up expired domains that were listed within the dataset, and hosting their own malicious data in its stead, they were able to poison any models created from this dataset. What's more they were able to poison 0.01% of the LAION-400M or COYO-700M datasets, for as little as $60. The researchers also discussed the possibility of poisoning up to 6.5% of Wikipedia by exploiting rolling snapshots on the site and timing their edits of pages accordingly.

What's concerning about these types of attacks is that you, or the creators of the model you're using, may be blissfully unaware that the data was poisoned to begin with, leading to potentially catastrophic downstream incidents.

To learn more about supply chain attacks within the context of AI applications, check out the blog Insane in the Supply Chain.

Threat Actors and Attack Vectors

Attacks on AI systems are already taking place in the wild, but the real scale to which they happen is difficult to assess. This attack vector is still very new, meaning that there is not enough awareness about it. As a result, **security solutions that could detect such attacks are few and far between.**

Model hijacking attacks, in which AI models are used to deliver traditional malicious payloads, are the easiest ones to spot. This is because existing software security concepts can be extended to detect and prevent such attacks.

They are also, from the attacker's perspective, the easiest ones to perform. The widespread lack of digital signing, integrity checking, and anti-virus scanning of AI artifacts makes them an enticing target for traditional cybercrime. Many security researchers have been subverting ML models to achieve code execution for proof-of-concept purposes. But it's not just security researchers that are looking into this attack vector. Several instances of hijacked models can likely be attributed to malicious actors. This includes models containing reverse-shells, as well as CobaltStrike and Metasploit stagers, all of which were connected to known malicious command and control centers.

Because hijacked models are often uploaded to public repositories, there is some visibility into them. However, the situation gets much more complicated with data poisoning, model evasion, and model theft attacks. **Most businesses do not monitor their AI for adversarial inputs**.

Those who do are not obliged to disclose that they've noticed malicious activity. Therefore, the details of adversarial attacks are rarely made public. Whatever is disclosed is most likely just a tiny tip of an iceberg - and the iceberg is poised to grow exponentially over the coming years, as more and more adversaries target AI systems.

The scarcity of information means it is too early to have a solid insight on threat intelligence regarding attacks on AI systems. However, it's definitely a good time to initiate discussion around it, and start collecting and organizing data.

PART 3:

ADVANCEMENTS IN SECURITY FOR AI

Before anyone can start implementing protections for certain technologies, the industry needs to figure out the ways in which these technologies are vulnerable. This is why offensive security plays such a big role in planning the defenses.

When a new technology comes out, white-hat researchers try to get one step ahead of the attackers and come up with proof-of-concept scenarios for potential attacks against this technology. Defensive solutions are often built upon previous offensive research and attack tooling.

Security for AI is no different. The first research papers and tools in this field were also of the offensive kind. For quite some time, attacks against AI were mostly covered in academia papers, with exercises performed by security professionals. However, the last couple of years have marked a massive shift.

With AI-based systems being rapidly implemented across sectors, there has also been a substantial rise in intentionally harmful attacks. The need for defensive solutions is now front and center. From MITRE ATLAS knowledge base, to NIST AI Risk Management Framework, to various national and international policies and regulations, defensive measures are now being implemented to lay the groundwork for securing AI.

Offensive Security Tooling for AI

Offensive security tooling has been around for a long time, enabling red teams and pen testers to evaluate IT systems for possible weaknesses. Although initially designed with security in mind, these frameworks have proven increasingly useful to malicious actors, enabling them to perform attacks with ease while only requiring an abstract understanding of how the attack works under the hood.

Projects such as Metasploit, Cobalt Strike, and Empire are now as much associated with malicious activity as they are with red-teaming. The concept of offensive security has also made its way to the field of artificial intelligence, where AI security researchers have developed various tools to test their attack techniques.

There are many publicly available security evaluation tools designed to test AI systems.

They offer various attack techniques from bypass to theft to code execution. Although very valuable in improving the security, safety, and robustness of the models, they can also be used by adversaries in malicious activities and make attacking AI more straightforward and accessible than it might at first seem.

Automated Attack Frameworks

Adversarial ML Frameworks

One of the first libraries for testing the robustness of AI systems against adversarial examples, called CleverHans, dates as far back as 2016. In 2018, IBM released its Adversarial Robustness Toolbox (ART), a framework that implements a multitude of attacks against AI and includes easy-to-follow Jupyter Notebook examples. MLSploit, a user-friendly cloud-based framework whose name calls out to Metasploit, was released in 2019; it allows for the creation of attacks on various malware classifiers, intrusion detectors, and object detectors. In the same year, QData released TextAttack, a powerful model-agnostic NLP attack framework that can help perform adversarial text attacks, text augmentation, and model training.

2020 saw the release of Armory, a containerized testing tool for evaluating adversarial defenses, which interfaces with IBM's ART. In 2021 Facebook released AugLy, a data augmentation library, which can be potentially used to generate adversarial examples. Microsoft followed with the release of Counterfit, an easy-to-use command-line automation layer for security evaluation of ML models, which interfaces with existing attack tools and frameworks, including ART, TextAttack, and AugLy, and resembles Metasploit in terms of commands and navigation.

Anti-Malware Evasion Tooling

In addition to robustness evaluation frameworks, there are also more specialized tools that aim at a specific outcome. MalwareGym, for example, helps bypass AI-based anti-malware solutions. Released in 2017 anti-virus company Endgame, it implements reinforcement learning in the modification of Windows applications. By taking features from benign executables and adding them to malicious ones, MalwareGym can create malware that bypasses malware scanners. Although MalwareGym is just a demo tool against one specific classifier, it has a successor in the MalwareRL project, which was released in 2021 and supports attacks against three different classifiers.

Model Theft Tooling

Another type of adversarial tool is KnockOffNets, released by researchers at the Max Planck Institute for Informatics in 2021. It's a tool for creating a replica of an AI model or, in other words, for stealing the model. It requires no previous knowledge of the model or the training data. The authors claim it can relatively accurately reproduce a model for $30. Although KnockOffNets was created to showcase the ease of model theft/model extraction attacks, it can also help adversaries build their own model theft tooling.

Model Deserialization Exploitation

Fickling, released in 2021 by Trail of Bits, is the first tool to exploit one of the most popular AI model serialization formats: Python's pickle format. It contains a decompiler, static analyzer, and bytecode rewriter and can inject arbitrary code into AI models saved as pickles. Charcuterie, released in 2022, implements a set of attacks that utilize code execution techniques and deserialization exploits in ML models.

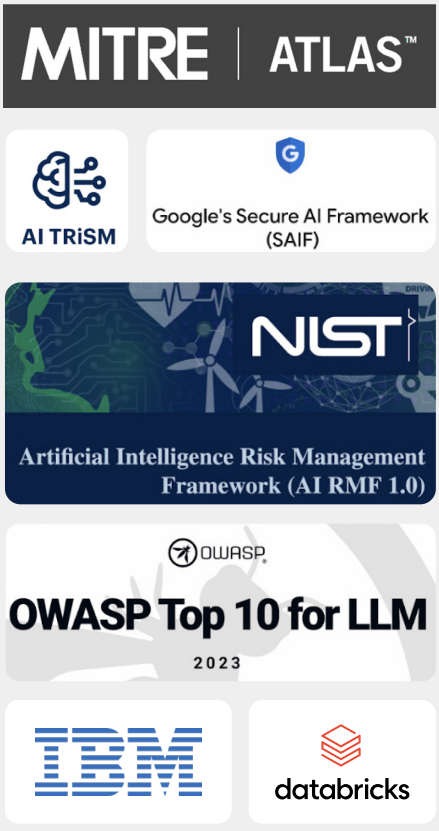
It was developed to demonstrate the vulnerability of ML models against more traditional attacks, but can be used by script kiddies to subvert publicly available models.

Defensive Frameworks for AI

With new tools and techniques for attacking AI popping up with increasing frequency, it has become clear that a methodical defensive approach is needed to safeguard this booming technology.

Over the last two years, several big cybersecurity players have created comprehensive frameworks comprising various security practices, strategies, and recommendations for AI. These frameworks are incredibly valuable first steps on the road long ahead.

Defensive Frameworks



MITRE ATLAS

First released in 2020 on GitHub then launched as a full website in 2021, MITRE ATLAS stands for "Adversarial Threat Landscape for Artificial-Intelligence Systems." ATLAS is a knowledge base of adversarial machine learning tactics, techniques, and case studies designed to help cybersecurity professionals, data scientists, and their companies stay up to date on the latest attacks and defenses against adversarial machine learning. The ATLAS matrix is modeled after and complementary to the MITRE ATT&CK framework, which is well-known and used in the cybersecurity industry to understand attack chains and adversary behaviors.

The ATLAS matrix is broken down into two main components: Tactics and Techniques. The tactics describe **what** an adversary is trying to accomplish. For example, Reconnaissance to learn more about a model deployment or Exfiltration to steal the model itself. The techniques, on the other hand, describe **how** an adversary is going to accomplish their tactic. Taking the Reconnaissance example, an attacker may Search Victim-Owned Websites for information about models or those internally who control or interact with them.

One difference between ATLAS and its ATT&CK counterpart is the source of the techniques. While ATT&CK is based only on detected attacks in the wild, ATLAS uses unique case studies selected from their impact to production AI-enabled systems. These case studies are a combination of both real-world attacks discovered in the wild as well as realistic red-teaming exercises from AI red teams or security groups. Some of these white hat hacker attacks are completely undetectable but have valuably demonstrated a realistic attack pathway that could threaten real-world AI-enabled systems. Including case studies of both malicious attacks and the white hat hacker attacks in ATLAS provides a more grounded and complete picture of the AI-enabled system threat landscape. These case studies outline who attack victims are and map to various techniques observed within the full scope of the attack.

> "This survey demonstrates the prominence of real-world threats on AI-enabled systems, with 77% of participating companies reporting breaches to their AI applications this year. The MITRE ATLAS community is dedicated to characterizing and mitigating these threats in a global alliance. We applaud our community collaborators who enhance our collective ability to anticipate, prevent, and mitigate risks to AI systems, including HiddenLayer and their latest threat report."
> – Dr. Christina Liaghati, MITRE ATLAS Lead

MITRE ATLAS Updates

In 2023, the ATLAS team released several major updates and new tools to continue enabling organizations that are working to secure their AI-enabled systems. These releases included:

- A significant update to the matrix to ground the rapidly evolving attack pathways for LLMs and GenAI enabled systems. This update added 12 new techniques and 5 unique case studies to ATLAS as the result of a close collaboration with Microsoft and other ATLAS community members determined to realistically represent these new LLM attack pathways.

- Arsenal and Almanac plugins developed collaboratively with Microsoft to add implementations of ATLAS techniques and new adversary profiles that target AI-enabled systems to CALDERA, an existing MITRE open-source threat emulation tool largely leveraged by the cyber world.

- An initial release of 20 new mitigations based on ATLAS case studies that provide high-level information on the security concepts and classes of technologies that can be used to prevent an adversarial attack technique from being successfully executed.

To date, ATLAS now has 14 Tactics, 82 Techniques, 22 Case Studies, and 20 Mitigations. As the ATLAS team continues to work with leading AI security organizations and experts across government and industry to expand the framework and its related tools and capabilities, the community-driven knowledge base and tools will remain a critical grounding resource as we all work to better secure our AI-enabled systems and supply chain against attacks.

In 2024, the MITRE ATLAS team will continue building upon the existing framework, tools and capabilities to help the community navigate the landscape of threats to AI-enabled systems by expanding on their platforms for both public vulnerability reporting and protected incident sharing. Through continued collaborations with industry, academia, and government, the ATLAS team is evolving open-source resources like the AI Risk Database, a tool for discovering vulnerabilities associated with public AI models. While the public ATLAS website continues to publicly represent unique real-world attacks, the ATLAS team is also continuing to expand its platform for more rapid protected or anonymized threat sharing within its community.

NIST AI Risk Management Framework

In January 2023, US National Institute of Standards and Technology released the AI Risk Management Framework (AI RMF). The AI RMF is a conceptual framework that takes learnings from the traditional software and information-based systems and applies them to the unique challenges presented by AI systems. It provides guidance for responsible design, development, deployment, and use of AI systems to give organizations additional trust in AI.

The framework splits itself into two parts: framing the risks related to AI systems and the core framework itself. The core describes four functions: govern, map, measure, and manage. Each breaks down into further controls to give organizations greater insights securing their AI infrastructure.

Google Secure AI Framework

Introduced by Google in June 2023, Secure AI Framework (SAIF) is a conceptual framework that, like NIST AI RMF, provides guidance on securing AI systems. It builds upon best practices and experience from traditional software development, adapting them to fit the needs of AI systems.

There are six core elements to the framework:

- Expand strong security foundations to the AI ecosystem
- Extend detection and response to AI
- Automate defenses to keep pace with existing and new threats
- Harmonize platform level controls
- Adapt controls and mitigations for AI deployment
- Contextualize AI risks to match business processes.

As with many other frameworks, SAIF will review traditional security controls around data and network level access, AI/ML specific controls such as data poisoning and detecting anomalies, privacy requirements and regulations, as well as governance around the entire lifecycle.

OWASP Top 10

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.

In 2023, OWASP released the Top 10 Machine Learning risks. These controls help those who are building, operating, and securing machine learning to identify potential risks and attack vectors within their deployments. Each of the individual controls has information on the attack vector, various risk factors that can help with threat modeling, and guidance on how to prevent the attack. When combined with other practical guidance from other frameworks such as ATLAS, this helps demystify the real threats to machine learning and what can be done about them.

Another recent release from OWASP are the top 10 critical vulnerabilities seen in Large Language Models (LLMs). With the rapid recent adoption of LLM technology, risks associated with deploying LLMs have been proliferating (as discussed in section 2). This OWASP document covers items such as prompt injection, output handling, all the way to model theft of the LLM itself. Each section also offers practical guidance for using this technology in a responsible and secure manner.

Gartner AI Trust, Risk, and Security Management (AI TRiSM)

Gartner defined a framework to address concerns around AI and ML systems, called AI TRiSM. It covers challenges such as bias, privacy, and explainability while also touching on the security and risks of such systems. This provides a roadmap for organizations to build AI/ML systems that maintain trust, are reliable and fair, and secure by design.

Databricks AI Security Framework (DAISF)

The DAISF framework adopts a comprehensive strategy to mitigate cyber risks in AI systems. It provides insights into how ML impacts system security and how to apply security engineering principles to AI systems. It also offers a detailed guide for understanding the security and compliance of specific ML systems.

Actionable defense recommendations apply to 12 foundational components of a generic data-centric AI system:

- raw data
- data preparation
- datasets
- data and AI governance
- machine learning algorithms
- evaluation
- machine learning models
- model management
- model serving and inference
- inference response
- machine learning operations
- data and AI platform security

Within these components, Databricks identified 54 technical security risks. Their recommendations are based on the real-world evidence that adversaries compromise unsecured AI systems using simple tactics.

IBM Framework for Securing Generative AI

In January 2024, IBM released their Framework for Securing Generative AI, focused on the use of LLMs and other GenAI solutions in businesses and organizations. It provides defensive approaches by helping to estimate the most likely attack that can occur at each stage of the pipeline, and suggesting relevant safeguards and defenses.

IBM's framework consists of five steps:

- **Securing the data**: describes risks related to the data collection and processing phase, such as mishandling PII and privacy concerns
- **Securing the model**: deals with attacks that can occur during model development and training, including supply chain attacks, API attacks, and LLM exploitation;
- **Securing the usage**: relates to the live use of model in production and covers inference attacks, including prompt injection and model theft
- **Securing the infrastructure**: tapping into existing expertise to optimize and harden network security, access control, data encryption, and intrusion detection and prevention
- **Establishing governance**: putting guardrails in place that ensures AI systems don't stray from what they are intended to do and act as expected.

Red Teaming and Risk Assessment

First ideas of AI red teaming emerged in the late 2010s. At that point, AI systems were already known for their vulnerability to things like bias, adversarial examples, and general abuse. Even though, now, there's widespread acceptance that AI will define this decade, it's still mostly the major players - such as Google, NVIDIA, or Microsoft - who invest in building their own internally-focused teams dedicated to pentesting the AI solutions they develop and implement.

It would be unfair to mention these companies by name and not highlight some of the incredible work they have done to bring light to the security of AI systems:

- In December 2021, Microsoft published their Best practices for AI security risk management
- In June 2023, NVIDIA introduced their red team to the world alongside the framework they use as the foundation for their assessments
- In July 2023, Google announced their own AI red team following the release of their Secure AI Framework (SAIF).

14% of IT leaders say their company are planning and testing for adversarial attacks on AI models

Policies and Regulations

We've already discussed how AI is a double-edged sword: it can be easily turned against people, businesses, and societies, with far-reaching consequences that could prove devastating. For this reason, **it is imperative for governments around the world to introduce tight regulations on how AI can be used safely, legally, and ethically.**

The first regulations around the use of AI were implemented as part of the European Union's General Data Protection Regulation (GDPR). These were very limited in scope and related mainly to the need for certain AI systems to be explainable. An AI model is explainable only if it's possible for us, humans, to assess why the model returned a specific prediction. This is important in all applications that make critical decisions, or decisions that can have an impact on people.

In 2019, the Organization for Economic Co-operation and Development (OECD) adopted the Recommendation on Artificial Intelligence (the "OECD AI Principles"). It describes five principles and five recommendations for OECD countries and adhering partner economies to promote responsible and trustworthy AI policies.

In 2022, the EU proposed a more comprehensive AI Act that groups AI solutions into three categories: low-risk applications that have to adhere to transparency laws but are otherwise unregulated; high-risk applications that are subject to strict limitations; and applications that are deemed dangerous and are outright banned. A provisional agreement between the EU Council presidency and the European Parliament was reached on this proposal in December 2023. It's expected to become law soon.

On a national level, several countries have started introducing AI-specific legislations. Singapore's Model AI Governance Framework, whose first edition dates back to 2019, consists of 11 AI ethics principles, including transparency, explainability, safety, security, data governance, and accountability. Canada's Digital Charter Implementation Act (Bill C-27), dated June 2022, encompasses the Artificial Intelligence and Data Act (AIDA) , which addresses the responsible adoption of AI. The UK is currently fleshing out its Artificial Intelligence (Regulation) Bill, whose purpose is to make provisions for the regulation of artificial intelligence.

In October 2022, the US introduced the Blueprint for an AI Bill of Rights, a set of suggestions and guidelines concerning the development and use of AI systems. A year after, in October 2023, the US White House issued an executive order on the safe, secure, and trustworthy development and use of artificial intelligence. The order sets standards for AI safety and security. It outlines risks AI systems pose, such as threat to human safety, detection of AI generated content, as well as securing the AI ecosystem. There are also orders on protecting privacy for citizens from data collection and storage to eliminate bias and discrimination in machine learning models. Finally, the order also aims to promote innovation and competition for those looking to advance and secure AI systems. The executive order mentions the NIST AI RMF multiple times, which will likely be a framework organizations can leverage to guide the secure development and deployment of AI systems.

PART 4:

PREDICTIONS AND RECOMMENDATIONS

It's always fun to dust off the crystal ball and try to predict future trends in cybersecurity. AI has been the dominant factor in many threat reports from traditional cybersecurity vendors this year. While most focus on generative AI, we take a broader look at the AI ecosystem and predict how it may be abused by cybercriminals, nation-states, and general bad actors over the coming year.

Predictions for the next 12 months

1. Data scientists will partner with security practitioners to secure their models

The cybersecurity industry has been in a technological arms race with adversaries for several decades, as each new advancement brings unique security concerns that require bespoke security solutions. However, AI/ML security has been overlooked in the data science world; rapid advances in AI and ML often lack even basic security controls. This has led to many vulnerabilities in libraries and tooling that have become pillars of AI software development. We expect this trend to reverse slightly over the coming year, as researchers work rapidly to uncover vulnerabilities and help shore up defenses in the open-source ML projects. The emerging collaboration between data scientists and cybersecurity specialists will boost the security of the whole AI ecosystem.

2. Supply chain attacks using ML artifacts will become much more common

Due to inherent insecurities in the machine learning tool chain, there are many low hanging fruits for cybercriminals to exploit. Threat actors are increasingly turning their sights towards MLOps platforms and tooling. Look for supply chain attacks to become more common as the year progresses, and not just for traditional initial compromise and lateral movement purposes. The often sensitive nature of ML models and the data they touch makes them very attractive to cybercriminals. Attackers will increasingly leverage vulnerabilities in MLOps platforms to poison training sets and exfiltrate sensitive data used at train or inference time to gain a competitive advantage or abuse AI systems.

3. There will be a significant increase in adversarial attacks against AI

Inversion attacks to infer training data or model details, inference attacks to generate bypasses/misclassifications, and, ultimately, model theft attacks will also become much more common. All these attack techniques will be driven by ever-expanding research into adversarial ML by academia and industry – which is being made available through easy-to-use open-source software. What was once a complex undertaking is – and will continue to become – increasingly simple for mere script kiddies to implement.

4. Threat actors will automate hacking efforts with LLMs

Generative AI is where we expect to find the most significant can of worms. Cybercriminals already use LLMs to enhance existing attacks, from authoring more realistic phishing emails to generating unique malware payloads on the fly and improving social engineering efforts. It's not a stretch to envisage threat actors harnessing LLMs to automate hacking efforts, perform reconnaissance, and supplement cybercrime-as-a-service over the coming year.

In addition, as LLMs evolve from text generation to multimodal systems capable of producing text, images, and audio, we expect a sharp increase in political activists and those trying to influence society using disinformation.

Another interesting development in the world of LLMs is RAG, Retrieval-Augmented Generation, which enhances the model with external sources of information or ground truths. RAG-empowered LLMs will be ripe for abuse by attackers, who will seek to leak sensitive information using carefully crafted prompts, especially if trained on corporate data.

5. Deepfakes will be increasingly used in scam and disinformation

Armed with powerful tools that can generate almost impeccable video and audio, adversaries are poised to become much more successful in their attempts at deceiving people, be it for the purpose of defrauding money, extracting sensitive information, or spreading fake news. The traditional scam scenario - in which the attacker sends a message pretending to be a relative who lost their phone and needs money - is now acquiring a whole new dimension. Instead of text-based messages, cybercriminals will be shifting to deepfake audio and video calls, and these can prove challenging not to fall for.

The bigger the digital footprint a person leaves behind, the more realistic a deepfake instance of this person can be. Naturally, public figures such as artists, influencers and politicians will be both the most enticing and vulnerable targets. However, it can take as little as just a few photos to create a deepfake convincing enough to trick a non savvy person into giving away money or information, and cybercriminals will look to cash in on low profile targets as well.

As the political climate deteriorates and tensions grow between nations, state-sponsored adversaries will use carefully curated deepfakes to steer public opinion, manipulate political campaigns and disturb elections. Conspiracy theories will have a wider reach and fake news will become increasingly difficult to disprove. Even if we find a way to reliably tell authentic videos from fakes it might not help limiting the damage. Once manipulated, people often refuse to acknowledge facts or accept the truth. The best solution to prevent deepfake-induced harm is to prevent the proliferation of deepfakes themselves.

6. AI attack surfaces will expand while more organizations use advanced tools to combat threats

It has never been easier to develop, use, and implement AI within organizations.. This rapid integration into established processes is introducing an ever-expanding novel attack surface that is not protected by conventional security controls. Businesses will experience many growing pains this coming year, where AI is exposed or configured insecurely, leading to data breaches, compromise, or worse.

On the flipside, we also expect to see more widespread adoption of AI security principles across organizations, and democratization of advanced methods of monitoring model behavior and model security evaluation, which have been typically reserved for major enterprises. As a result, many more organizations will be able to identify and take actions against adversarial attacks.

Securing Your AI: Getting Started

Understanding and implementing extensive security measures for AI is no longer a choice. It's a necessity. Too much is at risk for organizations, government, and society at large. Security must maintain pace with AI to allow innovation to flourish. That is why it is imperative to safeguard your most valuable assets, from development to operation and everything in between.

But how should you get started?

Let this guide be a starting point to securing your AI systems. Whether you're a developer, data scientist, or an IT professional, ensuring your AI systems are secure will empower you and your organization to navigate the future confidently.

93% of IT leaders say they have implemented security for AI protocols, but

58% aren't sure these protocols are keeping pace with evolving threats.

1. **Discovery and Asset Management**

   - ☐ Begin by identifying where AI is already used in your organization. What applications has your organization already purchased that use AI or have AI-enabled features?
   - ☐ Evaluate what AI may be under development by your organization. How many data scientists or data engineers roles are you employing? How many are you hiring? How many are consultants?
   - ☐ Understand what pretrained models from public repositories may already be in use. Do you know what websites offer pre-trained models? Do you understand the network/web traffic to these sites and who may have already downloaded these models?

2. **Risk Assessment and Threat Modeling**

   - ☐ Conduct a benefit assessment to identify the potential negative consequences associated with the AI systems if those models were to be compromised in any way.
   - ☐ Perform threat modeling to understand the potential vulnerabilities and attack vectors that could be exploited by malicious actors to complete your understanding of your organization's AI risk exposure

3. Data Security and Privacy

- ☐ Go beyond the typical implementation of encryption, access controls, and secure data storage practices to protect your AI model data. Those controls will not effectively protect the data in your models from theft, alteration, or other forms of attack. Evaluate and implement security solutions that are purpose-built to provide runtime protection for AI models. Look for solutions that can span the vast array of file types, model types, and also be agnostic to on-prem or cloud deployments.
- ☐ Embed into your 3rd-party risk process an evaluation of your vendors' security for their AI capabilities. Ask how your vendors incorporate security into their AI development lifecycle, including how they scan their models for data poisoning and malicious executables. Find out how they provide real-time/run time protection to detect and stop various forms of attacks against the AI capabilities embedded in the solutions you bought from them.

4. Model Robustness and Validation

- ☐ Regularly assess the robustness of AI models against adversarial attacks. This involves pen-testing the model's response to various attacks such as intentionally manipulated inputs.
- ☐ Implement model validation techniques to ensure the AI system behaves predictably and reliably in real-world scenarios. This will help minimize the risk of unintended consequences.

5. Secure Development Practices

- ☐ Incorporate security into your AI development lifecycle. Train your data scientists, data engineers, and developers on the various attack vectors associated with AI. Make sure to include how to minimize potential attack surface early in the security development lifecycle.
- ☐ Identify the AI security architecture required to be instrumented for the runtime protection of your AI when the models go into production use.

6. Continuous Monitoring and Incident Response

- ☐ Implement continuous monitoring mechanisms to detect anomalies and potential security incidents in real-time for your AI. Require your vendors to utilize AI in their solutions to alert you to attacks that could compromise your data or business processes.
- ☐ Develop a robust AI incident response plan to quickly and effectively address security breaches or anomalies. Regularly test and update the incident response plan to adapt to evolving AI threats.

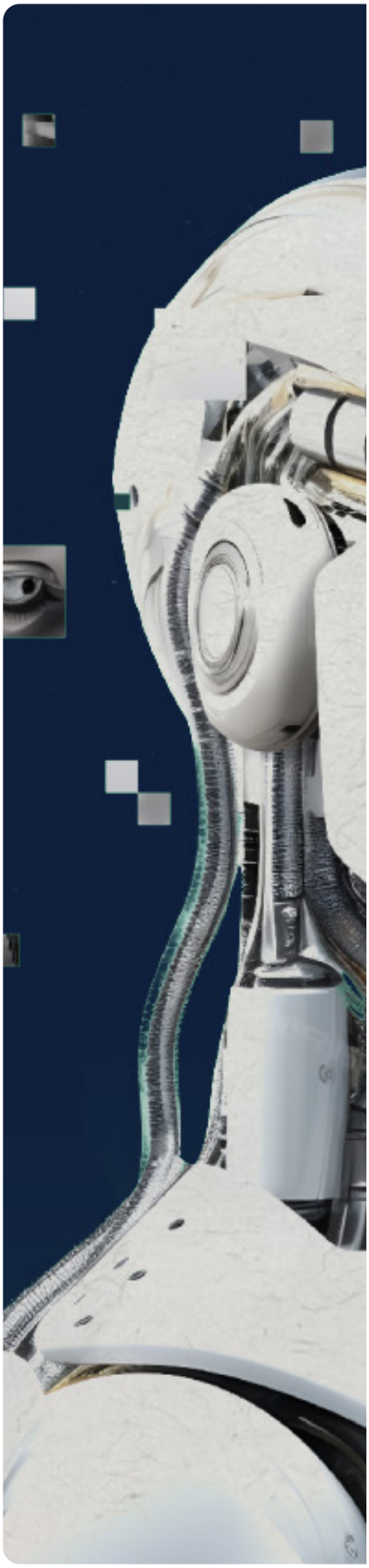
Remember that the security landscape – as well as AI technology – are dynamic and rapidly changing. It's crucial to stay informed about emerging threats and best practices. Regularly update and refine your AI-specific security program to address new challenges and vulnerabilities.

And a note of caution. **Responsible and ethical AI frameworks in many cases fall short of ensuring models are secure before they go into production, as well as 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.**

The final recommendation: Always ask yourself the following questions:

- What am I doing to secure my organization's use of AI?
- Is it enough?
- How do I know?

Only by answering these questions with data-driven, intellectual honesty, can you maintain the integrity of your security role and the critical function it provides.



Resources

HiddenLayer Products and Services

HiddenLayer AISec Platform

is a comprehensive AI security solution that ensures the integrity and safety of your models throughout the MLOps pipeline. By evaluating the security of pretrained models, detecting malicious injections, and monitoring algorithm inputs and outputs for potential abuse, the AISec Platform delivers an automated and scalable defense tailored for AI.

Learn More

HiddenLayer Machine Learning Detection & Response (MLDR)

complements your existing security stack, enabling you to automate and scale the protection of AI models and ensure their security in real-time. With MLDR integrated into your MLOps lifecycle and SIEM tools, you can proactively defend against threats to AI.

Learn More

HiddenLayer Model Scanner

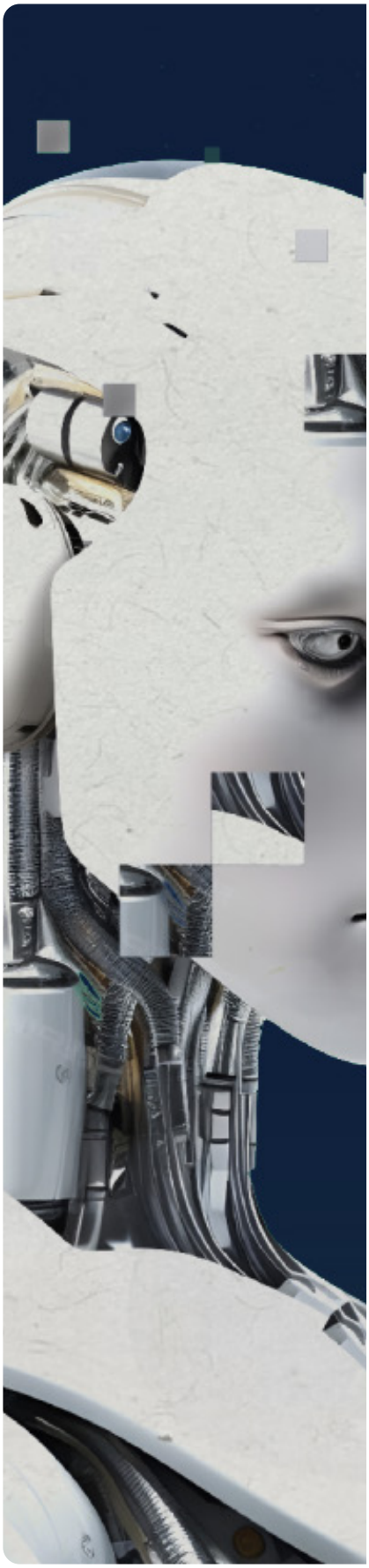
enables you to evaluate security and integrity of your ML artifacts before deploying them. This mitigates the risk of supply chain attacks through hijacked or backdoored models. With the Model Scanner, you can identify and remediate potential risks – ensuring a safe and trusted environment.

Learn More

HiddenLayer Professional Services

leverage deep domain expertise in cybersecurity and apply it to the field of AI. Our Adversarial Machine Learning Research (AMLR) team is equipped with a unique skill set that encompasses machine learning, reverse engineering, digital forensics and threat intelligence. We tailor our efforts to empower your data science and cybersecurity teams with the knowledge, insight, and tools needed to protect and maximize your AI investments.

Learn More



Get to Know HiddenLayer

AN INTRODUCTION TO HIDDENLAYER

Learn about HiddenLayer's origin story and what we are all about.

AN INTRODUCTION TO AI AND HOW TO PROTECT IT

Get a basic understanding of what Artificial Intelligence is and the pain points that exist in protecting it.

WHAT IS A MACHINE LEARNING MODEL

Dive deeper into what exactly a machine learning model is, and select use cases across industries.

AISEC PLATFORM OVERVIEW

Receive a high-level overview of HiddenLayer's AISec Platform. Learn more about what the platform provides as well as problems it helps solve.

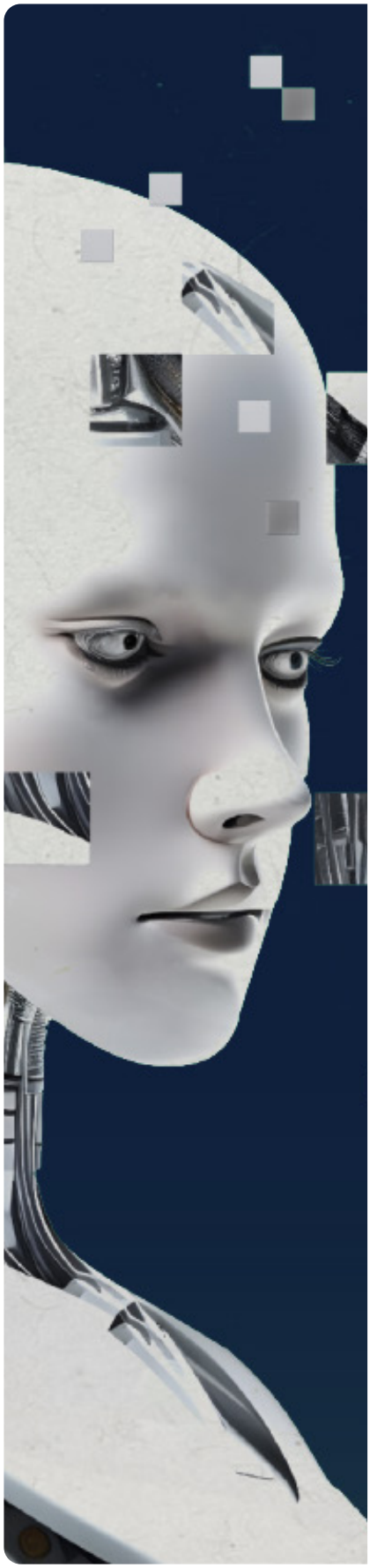
Read **GLOBAL FINANCIAL SERVICES CASE STUDY**

Explore a tangible customer case study that shows how HiddenLayer helped a top global financial services company minimize customer experience issues while combatting fraud.

HiddenLayer Research

HiddenLayer & Intel eBook: The Future of Risk is Upon Us
Companies can't adopt a zero-trust security posture without securing AI. Learn how to successfully navigate AI adoption and prevent malicious attacks.

Forrester Opportunity Snapshot: It's Time for Zero Trust
See how to take charge of AI security confidently, stay ahead of threat actors, and enable faster adoption of AI within your products and organization overall.



Securing AI: A Guide for SecOps

Read this comprehensive overview of the key considerations, risks, and best practices that should be taken into account when securing AI deployments within their organizations.

The Tactics and Techniques of Adversarial ML

Dive deeper into the details of adversarial attacks.

Weaponizing Machine Learning Models with Ransomware

See how easily an adversary can deploy malware through a pre-trained ML model with a destructive impact on an organization.

Insane in the Supply Chain

Understand the scope of your potential exposure through your supply chain risk management, as well as similarly affected technologies involved in machine learning and their varying levels of risk.

The Dark Side of Large Language Models

Learn more about the perils surrounding the use - and abuse - of generative AI.

Not So Clear: How MLOps Solutions Can Muddy the Waters of Your Supply Chain

This technical report publicly discloses six Zero-Day vulnerabilities in a well-known and widely used MLOps platform and demonstrates how the vulnerabilities can be combined to create a full attack chain against real-world systems.

Silent Sabotage: Hijacking Safetensors conversion on Hugging Face

Learn how an attacker could compromise the Hugging Face Safetensors conversion space and its associated service bot.

About HiddenLayer

HiddenLayer, a Gartner recognized AI Application Security company, provides security solutions for artificial intelligence algorithms, models, and the data that power them. With a first-of-its-kind, non-invasive software approach to observing and securing AI, HiddenLayer is helping to protect the world's most valuable technologies.

HiddenLayer was founded by AI professionals and security specialists with first-hand experience of how difficult adversarial AI attacks can be to detect and defend against. Determined to prove these attacks are preventable, the team developed a unique, patent-pending, productized AI solution to help all organizations protect important technology.

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, *VP of Research*
Andrew Davis, *Chief Data Scientist*
Sam Pearcy, *GTM Specialist*
Malcolm Harkins, *Chief Security & Trust Officer*
Travis Smith, *VP, ML Threat Operations*
Christina Liaghati, *PhD, AI Strategy Execution & Operations Manager at MITRE*

HIDDENLAYER
PROTECT YOUR ADVANTAGE