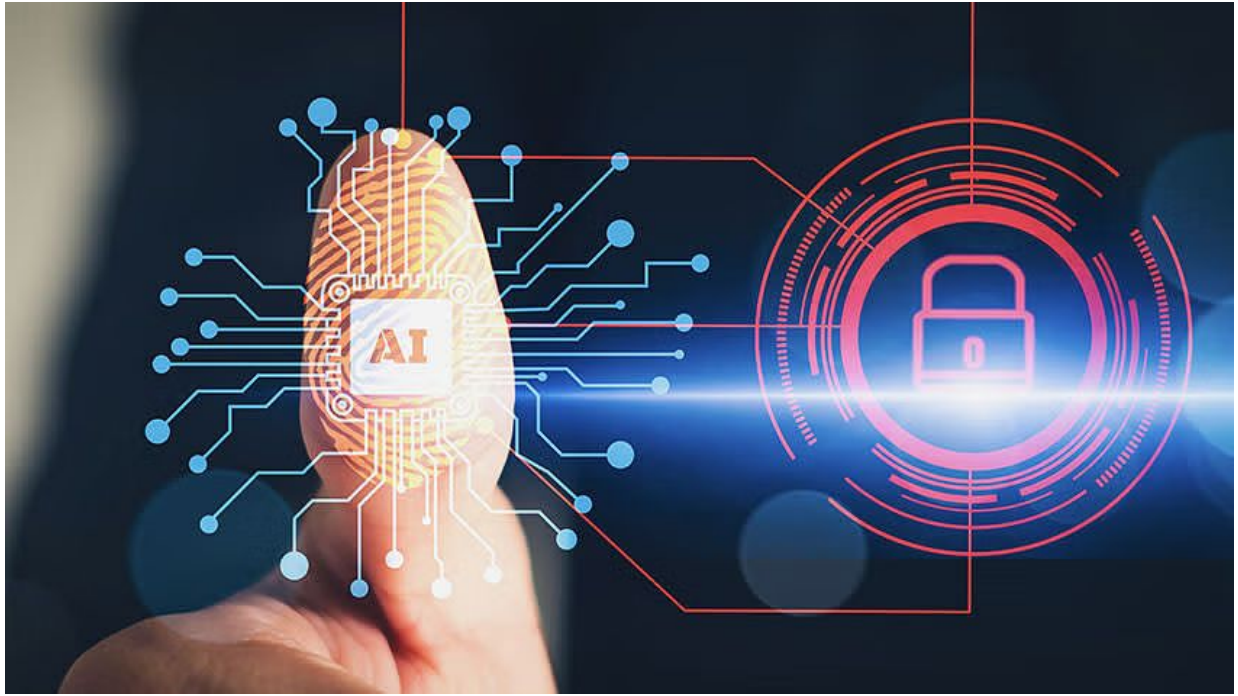
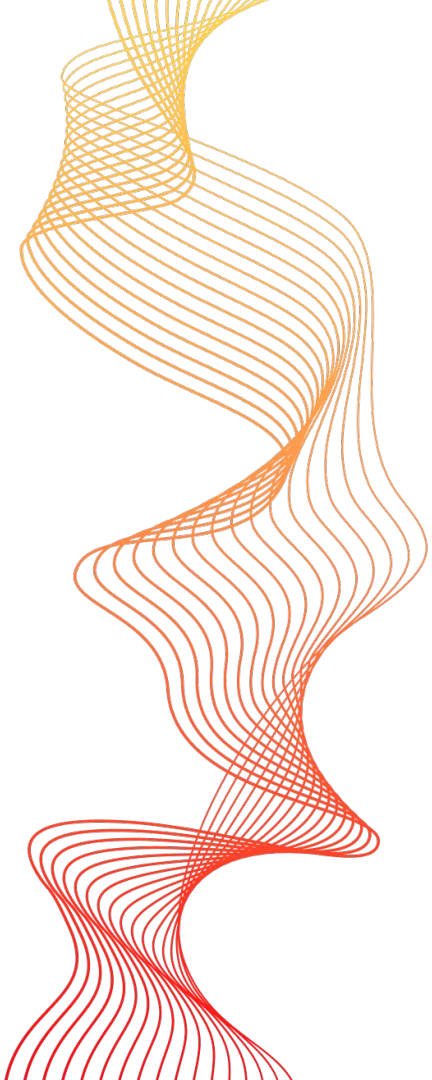
AI Security



Index



Why is AI Security Crucial?

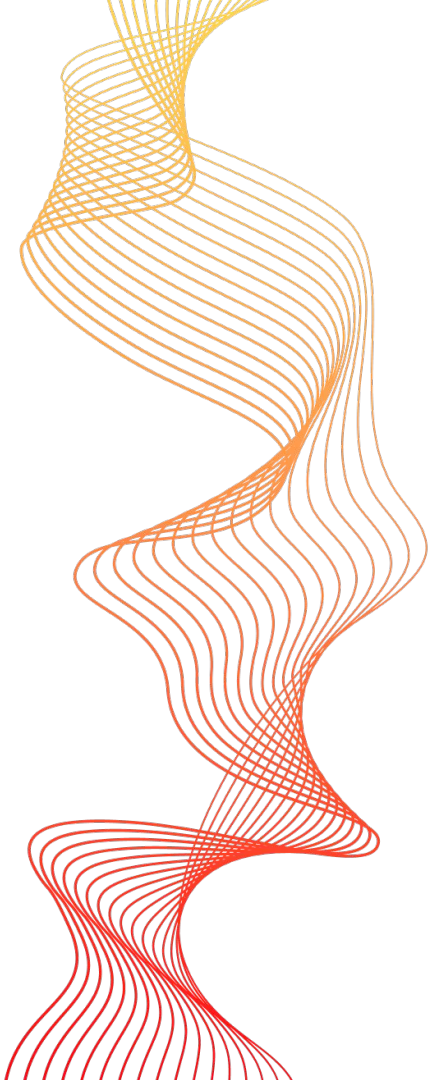
1. Data Privacy and Protection

2. Mitigating Malicious Attacks

3. Ensuring Reliability and Safety

4. Maintaining Trust and Integrity

Blog



AI Threat Landscape

Types of AI Threats

Adversarial Attacks

Model Stealing

Data Poisoning

Model Inversion

Evasion Attacks

Exploration Attacks

Blog

Adversarial Attacks

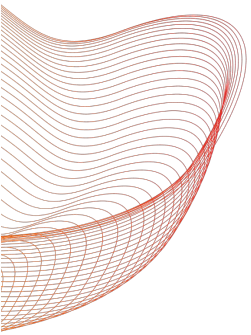
Definition: These attacks involve subtly manipulating the inputs to an AI system to cause it to make incorrect predictions or classifications.

Example: In 2020, researchers demonstrated that small, imperceptible changes to an image (such as altering a few pixels) could cause an AI image recognition system to misclassify a stop sign as a speed limit sign. This type of attack could have severe implications for autonomous driving systems, leading to potential safety hazards .

Data Poisoning

Definition: Data poisoning occurs when attackers inject malicious data into the training set, causing the AI system to learn incorrect patterns.

Example: In 2022, a group of attackers targeted a machine learning model used in a financial institution by introducing fraudulent transactions into the training data. This led to the model misclassifying legitimate transactions as fraudulent and vice versa, causing significant disruptions and financial losses .



Model Inversion

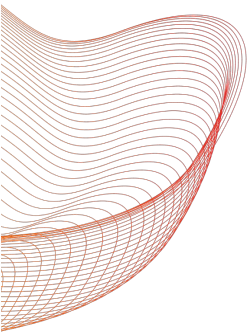
Definition: Model inversion attacks involve extracting sensitive information about the training data from the AI model.

Example: A 2021 study revealed that it is possible to reconstruct images of faces used in training a facial recognition system by exploiting the model's outputs. This poses significant privacy risks, especially if the training data contains sensitive personal information .

Model Stealing

Definition: Model stealing involves replicating an AI model by querying it extensively to understand its behavior and then recreating a similar model.

Example: In 2023, a cybersecurity firm demonstrated that they could replicate a proprietary natural language processing model deployed by a major tech company by sending numerous queries and analyzing the responses. This exposed the company’s intellectual property and could lead to competitive disadvantages .



Evasion Attacks

Definition: In evasion attacks, the attacker modifies the input data in such a way that the AI system is unable to correctly classify it, allowing malicious activities to go undetected.

Example: In 2022, cybersecurity researchers showed that malware could be slightly modified to evade detection by AI-based antivirus software. These modifications were minor enough to not alter the malware's functionality but significant enough to bypass the detection algorithms .

Exploration Attacks

Definition: These attacks aim to understand the AI system's inner workings by probing it with various inputs, potentially revealing weaknesses that can be exploited.

Example: In 2021, hackers conducted extensive probing of a speech recognition system to identify and exploit specific weaknesses that allowed them to bypass voice authentication mechanisms, highlighting vulnerabilities in biometric security systems .

What are attacks on AI?

Incorporating AI into a larger system can make the system susceptible to novel attacks that specifically target the AI. The techniques that adversaries use to carry out these attacks are distinct from traditional cyber techniques. By improving their understanding of these adversarial techniques, teams can work to mitigate the risks associated with AI incorporation.

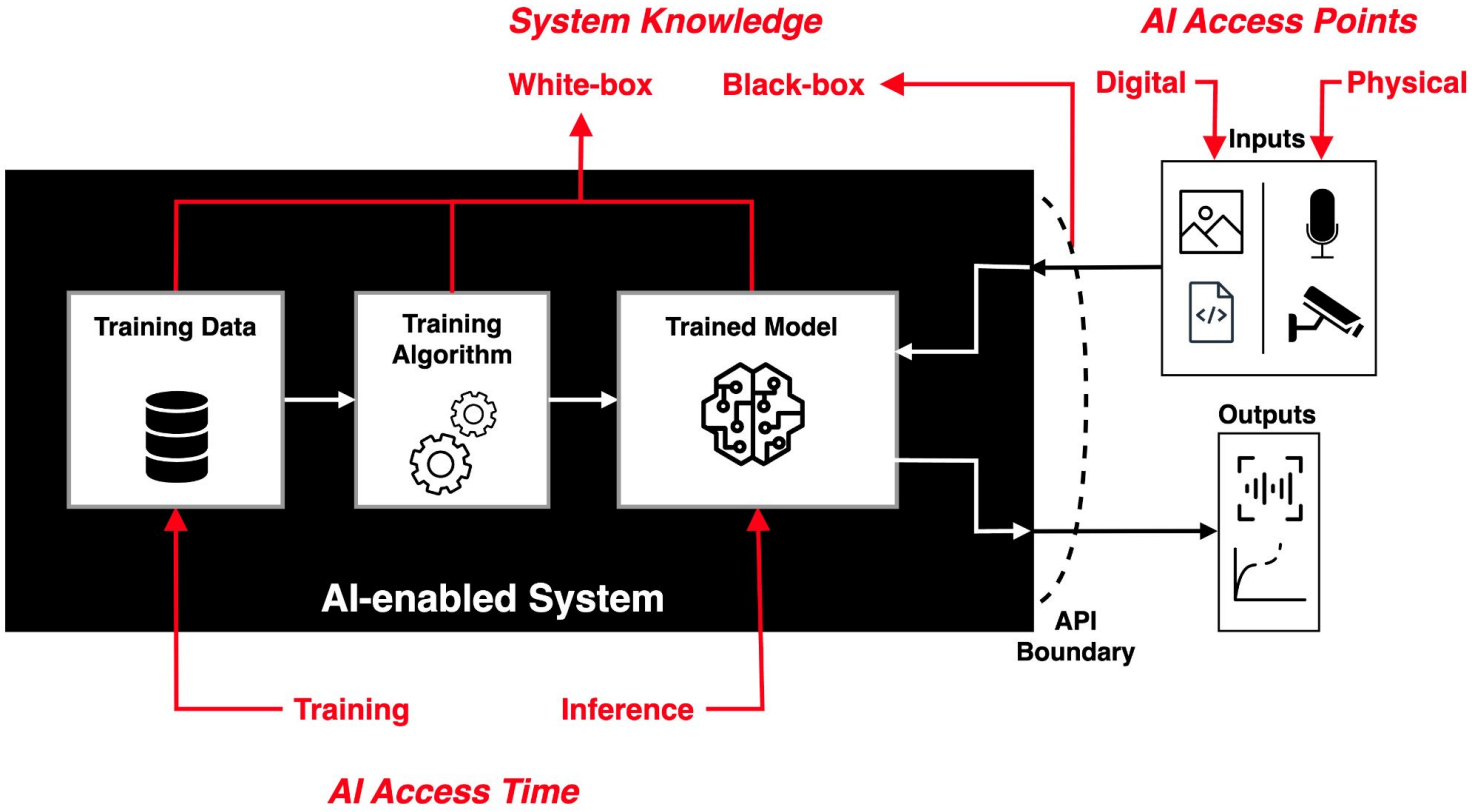
To better understand threats the wide range of effective attacks that can be used against to an AI-enabled system, we describe three important concepts that dictate an adversary's path of attack: AI Access Time, AI Access Points, and System Knowledge.

AI Access Time can be broken into two stages, ***training*** **and** ***inference*****.** The training stage is a process that includes collecting and processing data, training a model, and validating the model's performance. The end of the training stage and beginning of the inference stage occurs once a model is deployed. During the inference stage, users submit queries, and the model responds with predictions, classifications, or generative content known as the outputs (or inferences).

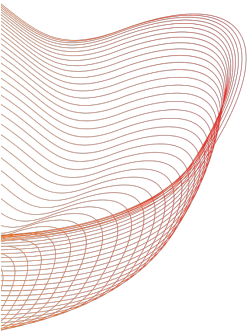
AI Access Points can either be *digital* or *physical*. A common digital access point within an AI-enabled system is API (application programing interface) access, where an adversary can interact with the model by sending a query and observing the response. A physical access point is used when an adversary interacts with data in the real world and influences the model's behavior by physically modifying the data collected.

System Knowledge refers to the amount of information an adversary knows about the ML components of the system. This knowledge can range from white-box, where adversaries have access to the model architecture, model weights and training data, to black-box where access and knowledge is limited to input and output responses during the inference stage (e.g. API access).

The figure below depicts an example of an AI-enabled system containing a trained AI model and the different types of access time, access points and system knowledge an adversary could leverage.


System Knowledge
AI Access Points
White-box
Black-box
Digital
Physical
Inputs
Training Data
Training Algorithm
Trained Model
Outputs
AI-enabled System
API Boundary
Training
Inference
AI Access Time

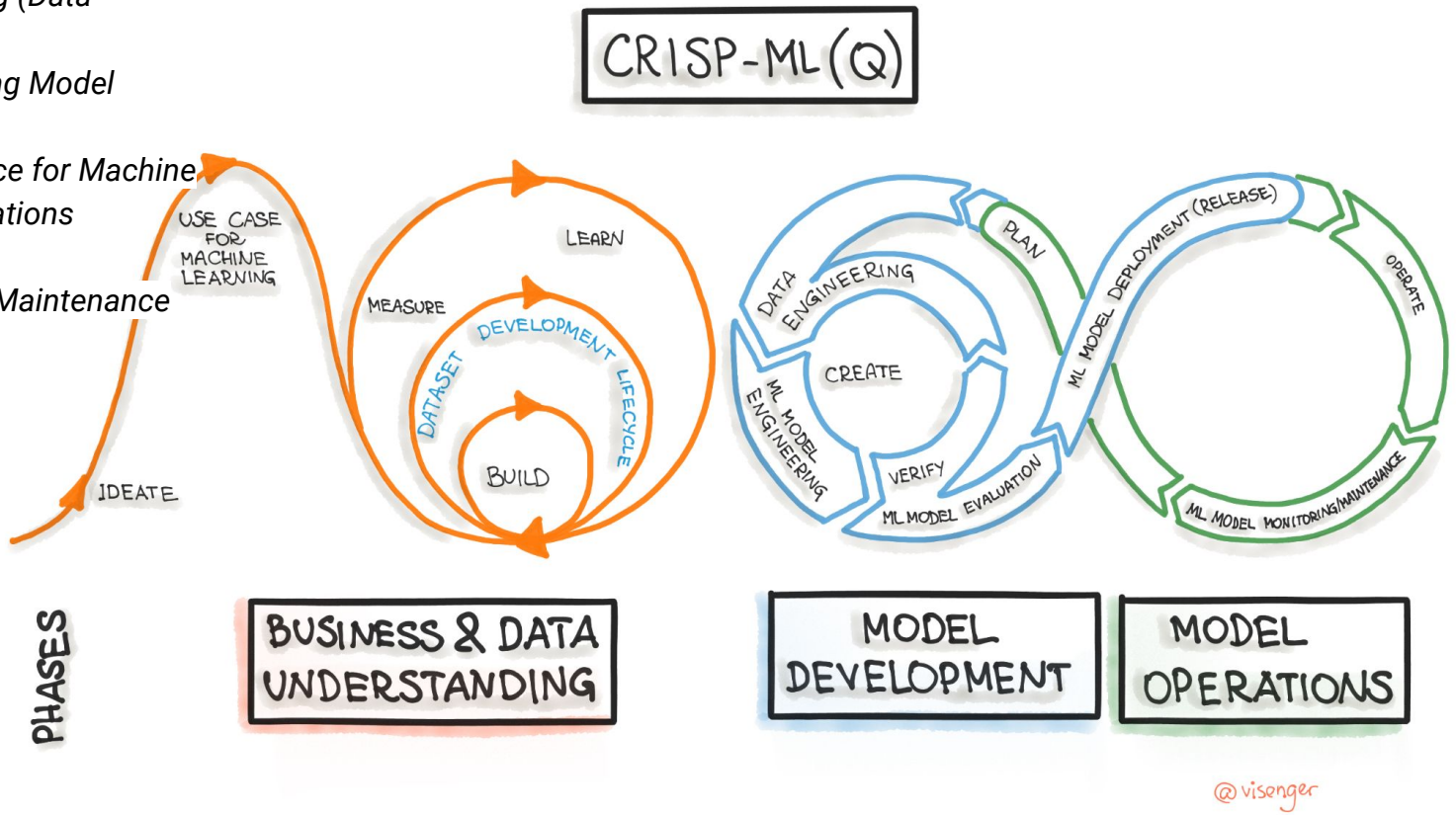
Attack	Overview
Poisoning Attack	Attacker modifies the training data of an AI system to get a desired outcome at inference time. With influence over training data, an attacker can create backdoors in the model where an input with the specified trigger will result in a particular output.
Evasion Attack	Attacker elicits an incorrect response from a model by crafting adversarial inputs. Typically, these inputs are designed to be indistinguishable from normal data. These attacks can be targeted, where the attacker tries to produce a specific classification, or untargeted, where they attempt to produce any incorrect classification.
Functional Extraction	Attacker recovers a functionally equivalent model by iteratively querying the model. This allows an attacker to examine the offline copy of the model before further attacking the online model.
Inversion Attack	Attacker recovers sensitive information about the training data. This can include full reconstructions of the data, or attributes or properties of the data. This can be a successful attack on its own or can be used to perform other attacks such as Model Evasion.
Prompt Injection Attack	Attacker crafts malicious prompts as inputs to a large language model (LLM) that cause the LLM to act in unintended ways. These "prompt injections" are often designed to cause the model to ignore aspects of its original instructions and follow the adversary's instructions instead.
Traditional Cyber Attack	Attacker uses well-established Tactics, Techniques, and Procedures (TTPs) from the cyber domain to attain their goal. These attacks may target model artifacts, API keys, data servers, or other foundational aspects of AI compute infrastructure distinct from the model itself.



CRISP-ML(Q) defines six phases in the model lifecycle:

1. *Business and Data Understanding*
2. *Data Engineering (Data Preparation)*
3. *Machine Learning Model Engineering*
4. *Quality Assurance for Machine Learning Applications*
5. *Deployment*
6. *Monitoring and Maintenance*

How does security fit into AI model lifecycles?



LLM Security

Large Language Models (LLMs) are a particular category of natural language models trained on hundreds of billions of words that can generate text or images and videos in response to natural language prompts. They vaulted to public popularity with the release of OpenAI's ChatGPT in November of 2022 due to their ability to perform multiple complex tasks such as content generation, style transfer, and text summarization, all with a single model.

From a security perspective, these systems introduce unique challenges to an AI pipeline due to the massive size of the training dataset, opaque internal architecture of the model, and use of natural language for input prompting. For example, indirect prompt injection attacks can be used to extract a user's personally identifiable information (PII) or influence the user to visit malicious websites.

For sample adversarial techniques, see LLM Prompt Injection, Compromise LLM Plugins, and LLM Jailbreak.

Implications for AI Security

Increased Vigilance: Organizations must stay ahead of attackers by continuously monitoring and updating their AI security measures.

Robust Training Data Management: Ensuring the integrity and quality of training data is crucial to prevent data poisoning attacks.

Model Privacy and Confidentiality: Techniques like differential privacy can help protect sensitive information used in training AI models.

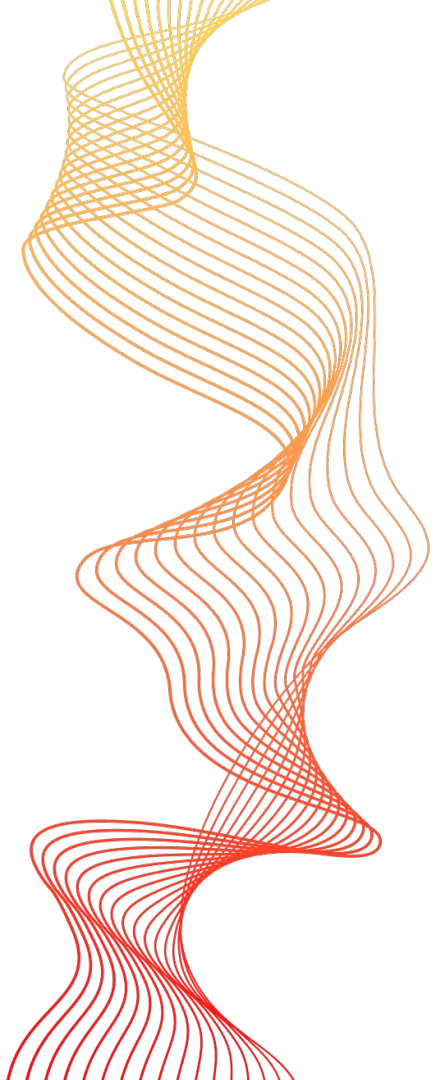
Enhanced Detection and Response: Implementing advanced detection mechanisms and robust incident response plans is essential to mitigate the impact of adversarial and evasion attacks.

Collaboration and Research: Ongoing research and collaboration within the cybersecurity and AI communities are vital to developing new defenses against emerging threats.

AI Security Best Practices

1. Data Security
2. Model Security
3. Operational Security
4. Development Security
5. Collaboration and Awareness

Blog



Thanks