AI Security 101

What is AI Security?

Artificial intelligence (AI) technology is advancing at a rapid rate and adoption is on the rise. Once limited only to highly controlled operational environments and use cases, today we see *AI-enabled systems* — software systems with one or more AI components — effectively integrated into a variety of use cases and available to the public.

AI security can be defined as the tools, strategies, and processes implemented that identify and prevent threats and attacks that could compromise the confidentiality, integrity, or availability of an AI model or AI-enabled system. AI security is a critical component of the AI development cycle to ensure safe and consistent performance throughout operation. In additional to the existence of traditional cybersecurity vulnerabilities, incorporating AI into systems also introduces new threat vectors and vulnerabilities that require a new set of security procedures. Identifying and mitigating these AI-enabled system vulnerabilities is an integral part of AI security and requires a technical and operational response.

In this 101, we describe common threats to AI-enabled systems documented within MITRE ATLAS™, security and the AI lifecycle, and active research areas.

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.

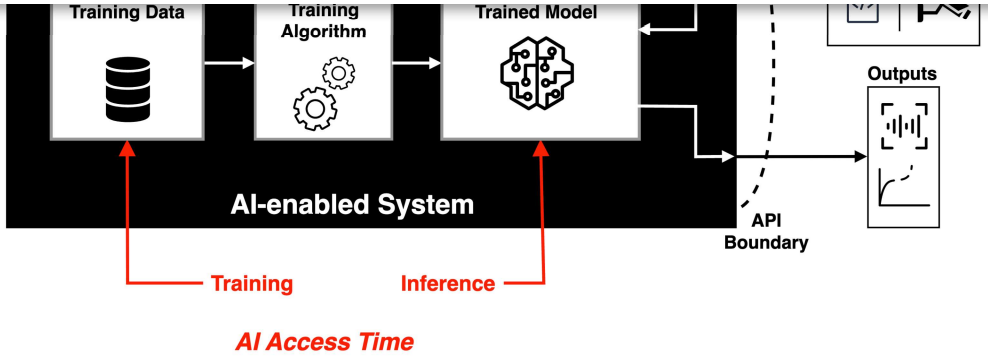

Figure 1: An AI-enabled system and key concepts.

The table below provides high-level descriptions of adversarial attacks and their possible effects on AI-enabled systems. For a comprehensive list we recommend exploring the ATLAS matrix.

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

	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.

How does security fit into AI model lifecycles?

An important consideration to countering attacks on AI-enabled systems is establishing clear operational procedures for managing a model throughout its lifecycle. Developing and deploying a robust AI model involves multiple phases of effort that typically involve different teams, developers, and stakeholders. Just as with the Software Development and Operations (DevOps) methodology, the field of Machine Learning Operations (MLOps) defines best practices and tools for deploying reliable, reproducible, and adaptable models. A good example of a model development pipeline with a MLOps focus is CRISP-ML(Q), the Cross-Industry Standard Process for the development of Machine Learning applications with Quality assurance.

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*

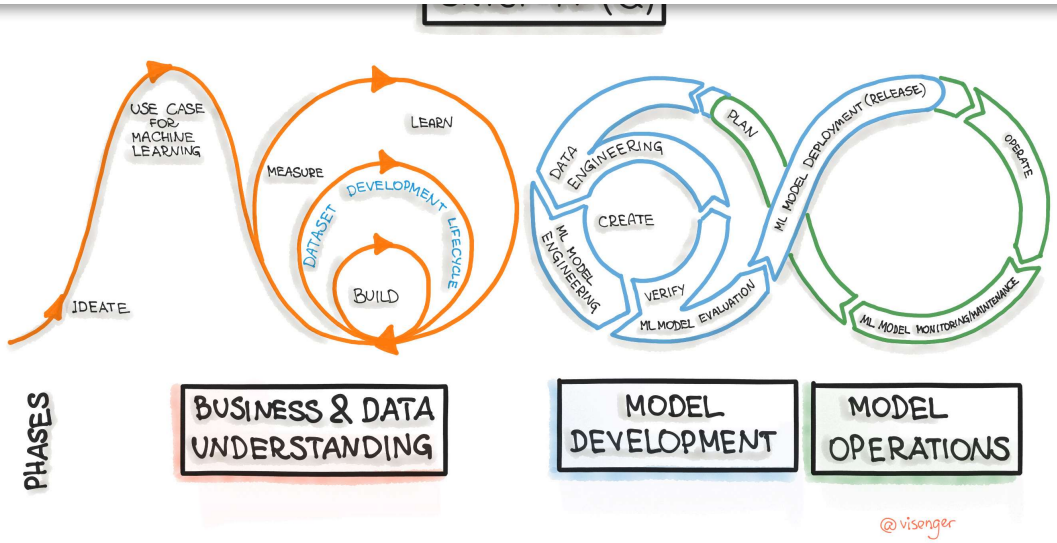

Figure 2: "The Machine Learning Development Life Cycle Process" by visenger. CC BY 4.0.

Each phase begins with defining the requirements and constraints of the task, then cycles through a process of risk identification, risk evaluation, and risk mitigation until requirements are met. Teams often revisit earlier phases and loop through the pipeline multiple times as stakeholders define new requirements and constraints.

It is expected that during the *Monitoring and Maintenance* phase that the process will return to earlier development phases in response to changing real-world conditions, such as concept drift, data drift, or malicious actions.

In ATLAS, we tag mitigations with phases from the CRISP-ML(Q) lifecycle to help each phase's teams identify vulnerabilities that could impact their task requirements and possible ways to respond. We also encourage interested parties to read the original paper on CRISP-ML(Q)

Recent AI Security Topics

AI Security is a constantly evolving field with subfields emerging as the technologies mature. We describe recent developments in three notable sub-fields below:

LLM Security

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.

We updated ATLAS in Fall 2023 to incorporate a new LLM focus that includes real-world case studies of adversarial attacks. In addition to this ATLAS work, we recommend this external list of LLM security related papers, articles, and tools for those interested in learning more.

Hardware Security

Hardware security has been studied extensively in classical cybersecurity settings and is now being examined in relation to AI systems. Example hardware security attacks include:

1. Side channel attacks — information about the system is deduced from alternative information streams such as voltage measurements or response timing,
2. Fault injection attacks — systems are actively disrupted by faulty input data or physical environment disruptions, and
3. Hardware Trojan attacks — malicious backdoors are inserted into the hardware of the systems including GPUs and other platform circuitry.

We refer interested readers to the following survey papers on this topic:

- Zhou et al. (2021): Deep Neural Network Security From a Hardware Perspective
- Xu et al. (2021): Security of Neural Networks from Hardware Perspective: A Survey and Beyond.

Policy

best way to balance interests of privacy and innovation in the United States, the Executive Order 14110 recently directed over 50 federal entities to take action across a range of AI policy areas. Several federal agencies have also enacted guidance on AI in recent years (e.g., GSA AI Guide for Government, DoD Responsible AI Strategy), but legislation over academic and private sector bodies remains a complex issue with technological, economic, and ethical considerations. The most effective way to balance these factors is an open research question.

We list a few leading relevant publications on this topic below:

- AI Risk Management Framework, NIST
- Guidelines for Secure AI System Development, UK National Cyber Security Centre
- A Unified Framework of Five Principles for AI in Society, Harvard Data Science Review
- A Taxonomy of Trustworthiness for Artificial Intelligence, UC Berkeley Center for Long-Term Cybersecurity
- Ethics and Governance of Artificial Intelligence for Health, World Health Organization
- A Sensible Regulatory Framework for AI Security, MITRE
- Strengthening and Democratizing the US Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource, National Artificial Intelligence Research Resource Task Force
- Artificial Intelligence Bill of Rights, Stanford University Human-Centered Artificial Intelligence
- S.3572 Algorithmic Accountability Act of 2022, 117th United States Congress
- EU Artificial Intelligence Act, European Parliament

Other recommended reading

- Our Tactics and Techniques pages list the methodologies adversaries use to infiltrate and/or compromise vulnerable AI systems. Our Matrix organizes these potential vulnerabilities graphically and chronologically for easier visual understanding, and our Mitigations page contains information how to protect your systems against them.

- Have suggestions on how we can make ATLAS more relevant to you and your organization? Contact us via Slack, LinkedIn, Email, or Github. We also welcome case study contributions through our interactive Case

Join our collaborative community to shape future tool and framework developments in AI security, threat mitigation, bias, privacy and other critical aspects of AI assurance.

CONNECT WITH US >

www.mitre.org