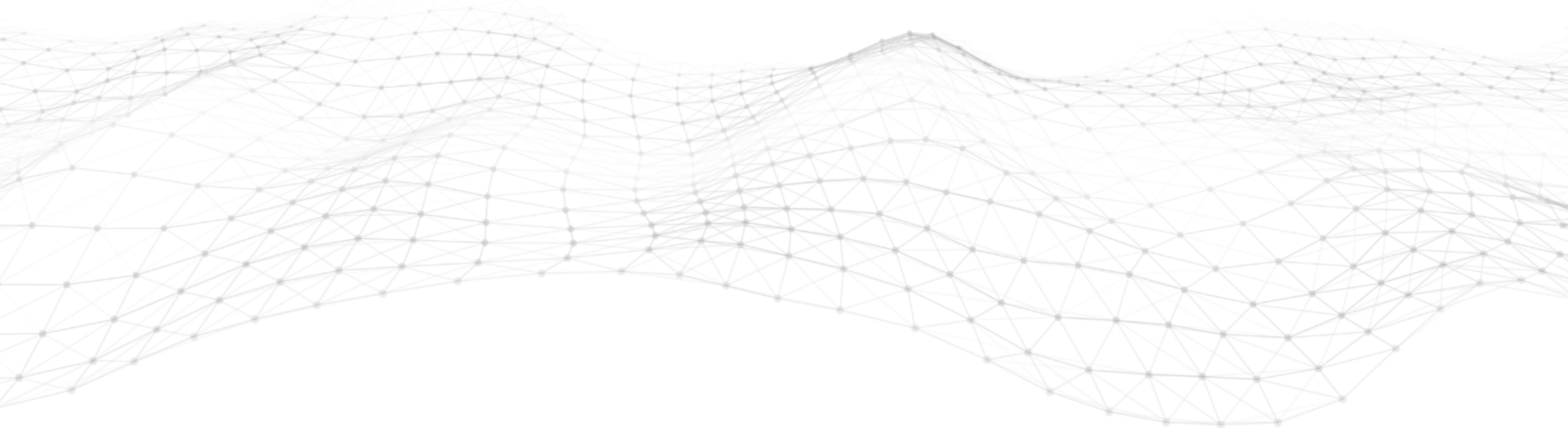
AI Security White Paper

09/26/2023

Come for the automation. Stay for the intelligence.



DeepSee™

Table of Contents

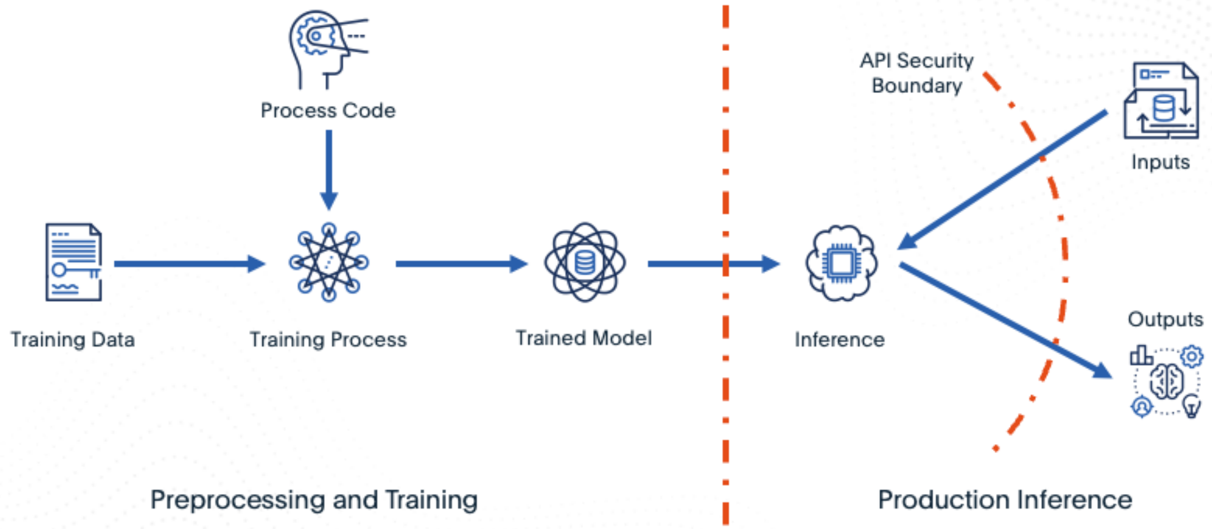
Introduction

This white paper provides an overview of the security challenges posed by AI and provides recommendations for organizations that are developing or deploying AI systems.

Background

Artificial intelligence (AI) is rapidly transforming the way we live and work. From healthcare to finance to transportation, AI is being used to automate tasks, improve decision-making, and create new products and services. We're at the beginning of a journey to destinations not yet imagined.

As AI becomes more pervasive, it is important to consider the security implications of this technology. AI systems are complex and often contain sensitive data, making them a target for cyberattacks. In addition, AI systems can be used to create new types of attacks, such as those that exploit the biases in AI models.



Threat Landscape

AI systems are vulnerable to intentional and unintentional threats, emanating from internal and external sources. Building and maintaining reliable AI systems require controlling inputs, close monitoring, and input conformance validation of model interactions. Predictions result from model training and validation data inputs and processing algorithms. Inputs have a direct effect on outputs, though it won't be immediately apparent. Observability of the processing pipeline, inputs, queries, and results should be monitored closely to identify and mitigate threat sources.

Internal Threats

Internal threats to AI models can come from a variety of sources, including:

- Malicious insiders: Employees who have access to AI models and data may be tempted to use this access for malicious purposes, such as stealing data or sabotaging the models.

- Human error: Even well-intentioned employees can make mistakes that can lead to security vulnerabilities, such as accidentally misconfiguring a system or leaving sensitive data exposed.

- Systemic weaknesses: AI models are complex systems that are often developed and maintained by multiple teams. This can lead to systemic weaknesses that can be exploited by attackers.

External Threats

External threats to AI models can come from a variety of sources, including:

- Hackers: Hackers may attempt to steal data from AI models or disrupt their operation.

- Competitors: Competitors may try to steal AI models or data, or they may try to disrupt the operation of AI models to gain a competitive advantage.

- Governments: Governments may attempt to restrict the use of AI models or to collect data from AI models for surveillance purposes.

AI Model Attacks

Known adversarial attacks on AI/ML systems typically follow one of these patterns:

Training attacks

- Poisoning attacks: In a poisoning attack, the attacker manipulates the training data to cause the model to learn incorrect patterns. This can be done by adding malicious data to the training set or by removing or modifying legitimate data.
- Backdoor attacks: In a backdoor attack, the attacker inserts a hidden trigger into the model that can be used to control the model's output. This can be done by adding malicious code to the training set or by modifying the model's architecture.

Inference attacks

- Evasion attacks: In an evasion attack, the attacker crafts inputs that are designed to fool the model into making incorrect predictions. This can be done by slightly modifying legitimate inputs or by generating adversarial examples.
- Adversarial attacks: In an adversarial attack, the attacker crafts inputs that are designed to cause the model to make a specific incorrect prediction. Adversarial attacks are often more difficult to defend against than evasion attacks because they are specifically designed to exploit the vulnerabilities of the model.

Other attacks

- Model extraction attacks: In a model extraction attack, the attacker attempts to steal the model from the victim. This can be done by exploiting vulnerabilities in the victim's system or by tricking the victim into downloading a malicious payload.
- Model inversion attacks: In a model inversion attack, the attacker attempts to reconstruct the training data from the model. This can be done by exploiting the mathematical properties of the model or by using adversarial examples.

Note that some attacks can be both training attacks and inference attacks. For example, a poisoning attack could be used to poison the training data of a model, which would then cause the model to make incorrect predictions at inference time.

This does not cover all types of attacks to ML systems; this is an active area of research and new attacks are being discovered regularly. Note that these attacks apply to ‘white box’ (known architecture, code, and training data) settings, and effective on any data.

Threat Protection

There are several threats that can impact AI systems, including:

- Data breaches: AI systems often require access to sensitive data, such as personal information or financial data. If this data is compromised, it could be used to harm individuals or businesses.

- Model manipulation: AI models can be manipulated by attackers to produce incorrect or biased results. This could lead to financial losses, damage to reputation, or even physical harm.

- System disruption: Attackers could disrupt the operation of AI systems, making them unavailable or unreliable. This could have a significant impact on businesses and organizations that rely on these systems.

To protect AI systems from these threats, organizations should implement a comprehensive threat protection program. This program should include the following elements:

- Data security: Organizations should implement strong data security measures to protect sensitive data from unauthorized access, use, disclosure, disruption, modification, or destruction.

- Model security: Organizations should implement security measures to protect AI models from manipulation or disruption. This could include techniques such as model encryption, access control, and auditing.

- System resilience: Organizations should implement measures to make their AI systems more resilient to cyberattacks. This could include techniques such as redundancy, failover, and disaster recovery.

By implementing a comprehensive threat protection program, organizations can help protect their AI systems from cyberattacks and mitigate the impact of these attacks.

In addition to the above, organizations should also consider the following best practices for threat protection:

- Keep AI systems up to date: Organizations should keep AI systems up to date with the latest security patches and updates.

- Monitor AI systems for suspicious activity: Organizations should monitor AI systems for suspicious activity, such as unusual patterns of behavior or errors in predictions.

- Plan for security incidents: Organizations should have a plan in place to respond to breaches in security and other incidents involving AI systems.

Governance, Risk, and Compliance

In addition to the technical security measures discussed above, organizations that are developing or deploying AI systems should also consider the governance, risk, and compliance (GRC) implications of this technology. GRC is a broad term that encompasses the processes and systems that organizations use to manage their risks and ensure compliance with applicable laws and regulations.

The GRC implications of AI can vary depending on the specific use case. However, some common GRC considerations include:

- Data privacy: Organizations that use AI systems to collect or process personal data must comply with applicable data privacy laws and regulations.
- Algorithmic bias: Organizations must be aware of the potential for algorithmic bias in AI systems and take steps to mitigate this bias.
- Explainability: Organizations must be able to explain how AI systems make decisions, so that users can understand how these decisions are made and trust the results.

By considering the GRC implications of AI, organizations can help to ensure that this technology is used in a responsible and ethical manner.

Additional Considerations

In addition to the security challenges and recommendations discussed above, there are several other considerations that organizations should keep in mind when developing or deploying AI systems. These considerations include:

- The role of humans in the AI lifecycle: AI systems are often designed to automate tasks or make decisions without human intervention. However, it is important to remember that humans still play a critical role in the AI lifecycle, from designing and building the systems to overseeing their operation and use.

- The impact of AI on society: AI has the potential to have a profound impact on society, both positive and negative. Organizations should be aware of the potential risks and benefits of AI and take steps to mitigate the risks while maximizing the benefits.

- The future of AI security: The security challenges posed by AI are constantly evolving. Organizations should stay up to date on the latest security threats and trends and implement security controls accordingly.

Recommendations

Organizations that are developing or deploying AI systems should take steps to mitigate the security risks associated with this technology. These steps include:

- Data security: Organizations should implement strong data security controls to protect the sensitive data that is used by AI systems. These controls should include encryption, access controls, and auditing, with all interactions mapped to a specific identity (persons and non-persons).

- Model security: Organizations should take steps to protect the security of AI models, such as by encrypting the models and storing them in a secure location. It’s paramount to understand, control, and log model inputs.

- Attack prevention: Organizations should implement security controls to prevent attacks on AI systems, such as by using input validation (semantic conformance) and anomaly detection.

- Response planning: Organizations should develop a response plan in case of an attack on an AI system. This plan should include steps for identifying and mitigating the impact of the attack. Model configuration management and version controls should be a requirement.

Conclusion

AI is a powerful technology that has the potential to transform many industries. However, it is important to be aware of the security risks associated with AI and to take steps to understand, identify, and mitigate these risks. By following the recommendations in this white paper, organizations can help protect their AI systems and data from cyberattacks.