Adversa AI expert feedback for the NIST AI RMF 2nd Draft

Feedback is provided by Eugene Neelou, Co-Founder & CTO at Adversa AI.

Adversa AI is on a mission to protect AI systems from cybersecurity threats.

Threat Modeling

Introduction

To attack AI systems, threat actors often target not only a variety of AI algorithms but also AI's software and hardware environments. This expands an attack surface increasing the number of possible attack vectors dramatically. Yet, it's not always feasible to protect AI systems from every possible attack.

Threat modeling activity is crucial to understand the most relevant threat actors and realistic attack scenarios. With the right threat model, AI stakeholders can prioritize their efforts in the design, development, and deployment of robust and secure AI systems.

Recommendation

Threat modeling should be individually listed in Subcategory in the MAP 1 Category.

It's fundamentally different from currently listed items, which in relation to other functions mostly set a context for intended AI functionality and unintended AI risks. However, such clauses barely include motivated threat actors willing to compromise an AI system.

MITRE ATLAS reference: https://atlas.mitre.org

Adversarial Testing

Introduction

To assess security posture, AI systems should be stress-tested with realistic AI attacks. Such adversarial testing should be conducted by experts in both machine learning and vulnerability exploitation following a valid methodology and probing common AI attacks as well as advanced state-of-the-art attacks.

Such a complex process needs a baseline to control adversarial testing coverage and quality. Based on the analysis of 5,000 academic works in adversarial machine learning, Adversa AI Red Team has created an unbiased list of standard AI vulnerabilities for AI security testing.

Recommendation

AI security and resilience in Subcategory MEASURE 2.7 should be extended.

Security testing should be based on a threat model covering all relevant attack scenarios across modification, infection, and exfiltration types. Comprehensive adversarial testing requires a variety of actionable security metrics. It's recommended to follow an AI security testing methodology and baselines such as the "Adversa Top 10 AI Vulnerabilities".

ADVERSA TOP TEN reference: https://adversa.ai/adversa-top-10-ai-vulnerabilities/

Security Lifecycle

Introduction

To protect AI systems from cyber threats, a change in AI development is required. It includes AI-focused integration of security practices into ML development and deployment workflow called MLSecOps as well as general cybersecurity management methods.

Security is unlike other listed AI risks. It focuses on threat actors motivated to compromise an AI system. The security of AI should have has its own lifecycle within the AI Risk Management Framework. There is an enormous amount of wisdom collected in the NIST Cybersecurity Framework. So, Adversa AI Red Team has applied the NIST CSF to AI security.

Recommendation

The MANAGE function should be extended in terms of Categories and Subcategories to support integration with current and future frameworks and methodologies.

NIST's AI Risk Management Framework should be synchronized with other NIST works such as the Cybersecurity Framework. Adversa AI Red Team has provided a sample lifecycle mapping NIST CSF to AI security. It enables improving MLSecOps workflows and integrating them into a wider cybersecurity management process.

ADVERSA SECURITY LIFECYCLE reference: https://adversa.ai/adversa-ai-security-lifecycle/

Adversa Top Ten AI Vulnerabilities

Adversa Top 10 is an unbiased and science-based list of the most important attacks.

1. **Evasion**
   The attack bypasses anticipated decisions by AI systems in favor of attacker-controlled behavior by crafting malicious data inputs.

2. **Poisoning**
   The attack reduces the quality of AI decisions while making AI systems unreliable or unusable by injecting malicious data into a dataset used for AI training.

3. **Membership Inference**
   The attack discloses characteristics of a specific data sample that was included in a dataset used for AI training.

4. **Backdooring**
   The attack enables hidden behavior of AI systems after poisoning training data with malicious triggers while AI systems work as expected in normal conditions.

5. **Model Extraction**
   The attack exposes algorithms' internal characteristics by making malicious queries.

6. **Attribute Inference**
   The attack reveals secret data characteristics by exploiting public information received from AI systems.

7. **Trojaning**
   The attack enables hidden behavior of AI systems after injecting or distributing malicious modifications of AI systems that work as expected in normal conditions.

8. **Model Inversion**
   The attack reveals secret inputs based on public outputs by making malicious queries.

9. **Resource Exhaustion**
   The attack makes AI systems use much more resources than in normal conditions leading to increased processing time, disrupted data flows, or denial of service.

10. **Reprogramming**
    The attack allows threat actors to repurpose AI systems and make them execute unintended tasks.