QuantUniversity, LLC

www.quantuniversity.com

NIST's Adversarial Machine Learning

Module 1: Introduction

1. Introduction

1.1 AI Systems and their Vulnerabilities

- **Artificial Intelligence (AI)** systems are being developed and deployed globally, leading to the emergence of AI-based services for people to use in various spheres of life.
- There are two main classes:
  - Predictive AI (PredAI)
  - Generative AI (GenAI)
- AI and machine learning (ML) technologies are vulnerable to attacks that may cause significant failures.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

1. Introduction

1.2 Gen AI

- GenAI is facing issues with **large language models (LLMs)** that are integral to the Internet infrastructure.
- They are being used for things like **online search, coding aid, powering chatbots**, and **Retrieval Augmented Generation (RAG)**.
- This new attack surface can expose confidential and proprietary enterprise data.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

1. Introduction

1.3 Privacy Concerns and Security Risks in AI Systems

- Companies developing AI models often do not release information about the used datasets. These unknown datasets may include sensitive personal information, such as addresses, emails, etc., creating a serious risk for **user privacy online**.
- The AI models' training data can be manipulated, making the AI systems vulnerable to attacks.
- Scraping of training data from the Internet also opens up the possibility of **data poisoning** at scale, leading to potential security breaches.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

1. Introduction

1.4 Privacy Concerns and Security Risks in AI Systems

- As ML models become more prominent, organizations often rely on pre-trained models, which could be adjusted with new datasets for different tasks.
- This process leads to opportunities for **malicious modifications of pre-trained models**, risking data leaks, incorrect processing, model availability, etc.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

1. Introduction

1.5 Contents of the course

- Standardized Terminology in AML
- Taxonomy of Attacks
  - Goals and Objectives
  - Attack Classes
  - Mitigations
- Mitigations in AML
- Pred AI Taxonomy
- GenAI Taxonomy
- Discussion and Remaining Challenges

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

QuantUniversity, LLC

www.quantuniversity.com

Thank you!

Contact

Email: info@qusandbox.com

www.QuantUniversity.com