Abdelhamid Mehri – Constantine 2 University
Faculty of New Information and Communication Technologies
Department of Fundamental Computer Science and Its Application

ETHICS AND GOVERNANCE OF ARTIFICIAL INTELLIGENCE

MASTER I SCIENCE DE DONNÉES ET INTELLIGENCE ARTIFICIELLE (SDIA)

DR ILHAM KITOUNI

2023-2024

LESSON8 : SECURITY IN AI

INTRODUCTION TO AI SECURITY

- The critical importance of security in the realm of Artificial Intelligence (AI).
- The capacity of AI systems to handle sensitive data and the need for robust security measures.

SIGNIFICANCE OF AI SECURITY

- **Data Protection:**
  - Safeguarding sensitive data, including personal and financial information.
- **Preventing Adversarial Attacks:**
  - Ensuring resilience against malicious attacks attempting to compromise AI systems.
- **Ensuring Reliability and Availability:**
  - Importance of maintaining the reliability and availability of AI systems for consistent performance.

POTENTIAL THREATS TO AI SECURITY

- **Data Attacks:**
  - Threats to the integrity and confidentiality of data used to train or operate AI systems.
- **Code-Based Attacks:**
  - Potential attacks targeting the source code of AI systems, allowing adversaries to gain control.
- **Infrastructure Attacks:**
  - Risks associated with attacks on the infrastructure supporting deployed AI systems, potentially rendering them inaccessible or non-functional.

EXAMPLE - ATTACK ON A FRAUD DETECTION SYSTEM

- **Scenario:**
  - How an AI system designed for detecting fraudulent activities in banking transactions could be targeted.
- **Potential Impact:**
  - Potential consequences, such as compromised credit card information.

TYPES OF ATTACKS

- **Data attacks**
  - Injection of fraudulent transactions into training data
  - Manipulation of existing data
- **Model attacks**
  - Exploiting adversarial examples
  - Poisoning the inference pipeline
- **Social engineering attacks**
  - Tricking users into revealing sensitive information
  - Bribing or blackmailing insiders
- **Technical vulnerabilities**
  - Exploiting software bugs
  - Denial-of-service attacks

HOW AI-POWERED BANKING FRAUD DETECTION SYSTEMS CAN BE TARGETED?

Developers and security professionals need to understand potential attack vectors to mitigate risks.

Some specific examples of attacks that could be used to target AI-powered banking fraud detection systems:

Data poisoning

- An attacker could inject fraudulent transactions into the training data for an AI fraud detection system. This could trick the system into learning to accept fraudulent patterns as legitimate.

Model manipulation

- An attacker could exploit adversarial examples to trick an AI fraud detection system into making incorrect predictions. In the context of fraud detection, an attacker could create transactions that are slightly different from real fraudulent transactions but are designed to trigger a false negative from the AI system.

HOW AI-POWERED BANKING FRAUD DETECTION SYSTEMS CAN BE TARGETED?

Social engineering

- An attacker could trick users into revealing sensitive information that could be used to bypass fraud detection controls. F
- or example, the attacker could send phishing emails or phone calls that appear to be from a legitimate source. Once the attacker has the user's credentials or account information, they can use them to initiate fraudulent transactions.

REFERENCES

- The case of the self-driving car that struck and killed a pedestrian : - "Uber self-driving car in fatal Arizona crash had software flaws: NTSB", Reuters, 2019. Disponible sur : https://www.cbc.ca/news/business/uber-self-driving-car-2018-fatal-crash-software-flaws-1.5349581
- The case of the AI system that was used to target people with advertising : - "Cambridge Analytica scandal: What you need to know", BBC News, 2018. Disponible sur : https://www.bbc.com/news/technology-43465968 and https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html