AI vs Security

IMPORTANT NOTES:
Study lecture materials at least 1 hour and prepare the questions prior to the tutorial session. The questions will be discussed in the tutorial session.

1. Define the terms ”AI for Security” and ”Security attacks AI”. Provide one real-world example for each.

   - AI for Security refers to using AI to enhance security measures, such as detecting anomalies or identifying threats. For example, AI-powered facial recognition for surveillance.
   - Security Attacks AI refers to exploiting vulnerabilities in AI systems to manipulate or deceive them. For example, poisoning a dataset to bias a hiring algorithm.

2. Explain the difference between conventional AI and robust AI in the context of security threats. Why is conventional AI considered ”too idealistic”?

   Conventional AI assumes datasets are benign and errors are accidental, making it vulnerable to malicious attacks. Robust AI is designed to withstand adversarial manipulation, such as poisoned datasets or coalition-based bias. Conventional AI is ”too idealistic” because it doesn’t account for intentional malice in data.

3. What are adversarial attacks in AI? How can they compromise the integrity of machine learning models?

   Adversarial attacks involve subtly altering input data (e.g., images, text) to trick AI models into making incorrect predictions. They compromise integrity by causing misclassification (e.g., making a stop sign unrecognizable to a self-driving car).

4. Discuss how collaborative multi-party AI (e.g., facial recognition across countries) could introduce bias into machine learning outcomes. What are the security implications?

   Collaborative multi-party AI (e.g., international facial recognition) may introduce bias if some parties contribute manipulated data. Security risks include:

   - Skewed models favoring certain groups.
   - Exploitation by malicious actors to evade detection.

5. Explain the concept of Generative Adversarial Networks (GANs). How do they relate to security in terms of both attack and defense?

   GANs (Generative Adversarial Networks) consist of two competing models: a generator (creates fake data) and a discriminator (detects fakes).

   - Attack: Used to create deepfakes or bypass detectors.
   - Defense: Improves fraud detection (e.g., spotting synthetic media).

6. In adversarial gaming (e.g., attacker vs. defender), why is the playing field often considered unfair? Compare this to AI vs human games like Chess or Go.

   In security, the attacker-defender dynamic is unfair because attackers need only one success but defenders must block all attempts. Unlike games (Chess or Go), defenders lack perfect information. Humans and machines have equal rules, but AI’s speed and memory gives it an edge.

7. A self-driving car’s AI misclassifies a stop sign due to an adversarial attack. What security goal is violated, and how could this be mitigated?

   Integrity is violated as the AI’s decision was corrupted. This can be mitigated as below:

- Adversarial training: The AI model is trained on both clean data and adversarial examples (e.g., stop signs with subtle perturbations). The model learns to recognize and resist malicious inputs, reducing misclassification.
- Robust model testing: Security teams simulate attacks (e.g., altering stop signs in test environments) to identify vulnerabilities before deployment. This reveals weaknesses in the model's decision boundaries, allowing fixes.
- Physical safeguard: Combine multiple sensors (cameras, LiDAR, radar) to cross-verify stop sign detection.

8. In a cybersecurity arms race, AI-powered malware evolves to bypass AI-powered defenses. Analyze this scenario using the adversarial gaming framework. Who has the upper hand, and why?

The scenario involves two AI-driven players with opposing goals:

- Attacker (AI-Powered Malware): Aims to infiltrate systems, evade detection, and cause harm (e.g., ransomware, data theft).
- Defender (AI-Powered Security): Aims to detect, block, and mitigate threats in real time.

This mirrors Generative Adversarial Networks (GANs), where the malware (like a generator) evolves to create attacks that fool defenses. The security system (like a discriminator) learns to flag malicious behavior.

Despite AI defenses, attackers typically lead due to inherent asymmetries:

- Attackers choose the time, target, and method. Defenders must react after an attack is detected.
- A failed attack can be refined and redeployed at near-zero cost. Defenders face high costs such as breaches lead to financial loss, reputational damage, and regulatory penalties.
- AI can generate infinite variants of malicious code to evade signature-based detection. For example, Malware can subtly alter its behavior to appear benign to AI detectors (e.g., mimicking normal network traffic).
- Defenders train on historical attack data, but attackers innovate new tactics. For example, AI-powered phishing emails now use LLMs (like ChatGPT) to craft highly personalized, convincing lures.