AICS
AI Hardening Slides

Maximillian See Tze Jie | 2102869A

Yee Zi Hung Anthony | 2102946G

Tan Guang Xuan Radcliffe | 2101135H

Poh Jie Ren Luke | 2102355A

Isaac Kwa Zheng Kai | 2104508C

Activity 1: Real-World Attacks on AI Systems

Use Case: Confusing Antimalware Neural Networks

Background of Attack

- ➢ Kaspersky researchers discovered a malware attack where the attackers manipulated malware files to evade detection by Kaspersky's ML model.
- ➢ The attackers used adversarial techniques to make subtle modifications to the malware files and confuse the ML model.
- ➢ The attack involved multiple stages and advanced techniques to compromise targeted systems.
- ➢ The malware was distributed through seemingly harmless files or documents in phishing emails or compromised websites.
- ➢ The files contained hidden malicious code that exploited vulnerabilities in popular software or operating systems.

Background of Attack Cont.

- Opening the infected file activated the malware's initial payload, leading to the establishment of a persistent presence on the compromised system.
- The malware employed complex obfuscation techniques and anti-analysis mechanisms to avoid detection by antivirus software and intrusion detection systems.
- Once the malware gained a foothold, it initiated further stages, including downloading additional modules and establishing communication with a command-and-control server.
- The additional modules facilitated malicious activities such as data exfiltration, privilege escalation, lateral movement within the network, and deploying additional payloads.

Those Affected

Alarming implications who rely heavily rely on such services, which include:

- ➢ Companies
- ➢ Users of Companies
- ➢ Governments
- ➢ Other unnamed critical infrastructure

How It Happened

- The features for models are built on users' systems and then sent to the cybersecurity company servers. The Kaspersky ML research team explored this gray-box scenario and showed that feature knowledge is enough for an adversarial attack on ML models.
- They attacked one of Kaspersky's anti malware ML models without white-box access to it and successfully evaded detection for most of the adversarially modified malware files.
- Malware files are modified in a way that evaded detection by the Kaspersky antimalware ML model. By manipulating the feature knowledge, which is essentially the information extracted from users' systems, the team crafted adversarial examples that the model failed to identify as malicious.

Impacts of Attack

- Researchers demonstrated that many adversarial files bypassed the antimalware model
- Attackers could deploy their custom-crafted malware, leading to:
  - Widespread system infections
  - Compromising national security
- For companies and governments:
  - Mistrust
  - Loss of reputation
  - Legal & Regulatory Issues
  - Loss of market share & Decrease in Sales

Mitigation Measures

- Advanced Threat Detection: Implement advanced threat detection solutions that go beyond signature-based detection methods. Utilize behavior-based analysis, machine learning algorithms, and anomaly detection to identify and block evasive malware.
- Security Courses for individuals
- Regular Updates and Patches: Keep all software, operating systems, and security solutions up to date with the latest patches and updates. This helps protect against known vulnerabilities that attackers may exploit to evade detection.

Activity 2: Attack Category of Aforementioned Use Case

Category: Evasion

Evasion Definition

- Creating inputs specifically designed to cause AI model to misbehave/fail to provide accurate predictions
- Thereby exploiting its vulnerability to alter small changes in input data to compromise security and reliability of model
- Can be executed by adding small disruptions to inputs (e.g. adding noise, altering inputs in a subtle manner

Activity 3: Real-World Incident of AI as Attack Tool

Use Case: Unusual CEO Fraud via Deepfake Audio Steals US$243,000 From UK Company

Background of Attack

- In 2019, an German energy company based in the United Kingdom fell victim to a deep fake voice scam. The attackers used AI-based voice synthesis technology to impersonate the firm's CEO, mimicking his voice and mannerisms.
- The attack consist of creating synthetic audio created using machine-learning algorithms to mimic the voice of the CEO and managed to convince an employee to transfer $220,000 to a supposed vendor.
- The fraudsters made subsequent calls to the targeted employee, but the company became suspicious and did not send the money after the second call. The suspicion arose because the second call came from an Austrian phone number instead of the expected German phone number associated with the CEO.

How It Happened

- To execute the attack, the attackers likely collected a substantial amount of voice data from publicly available sources, such as speeches, interviews, or other recordings featuring the CEO's voice.
- The attackers utilized deep learning algorithms and speech synthesis techniques to analyze and replicate the CEO's voice patterns and speech characteristics.
- They then generated a synthetic voice that closely resembled the CEO's tone & voice.
- Attackers contacted the energy firm's senior financial officer using the deep fake voice
  - To convince him to transfer a significant amount of money into a fraudulent account
  - According to one of the employees, the of yet unidentified fraudster called the company three times: the first to initiate the transfer, the second to falsely claim it had been reimbursed, and a third time seeking a follow up payment.

Impacts of Attack

- Successful scam resulted in:
  - Substantial financial loss for company
  - Mistrust & fear amongst employees concerning accounting processes in future
  - Loss of reputation amongst general public
- Operational disruption within the organization, and the following has to be dealt with:
  - investigations
  - legal proceedings
  - implementation improved security measures,

Mitigation Measures

- Implement robust authentication protocols/verification steps for accounting processes such as:
  - Multi-factor authentication (MFA)
  - Secure communication channels
- Training of employees to raise awareness about the existence of deepfake technology which:
  - Helps in verification of voice-based requests
  - & Is crucial in prevention of attacks involving sensitive transactions
- Long-term wise, to combat deep-fake technology
  - Special technology can be explored

Activity 4: Countermeasures

Countermeasure Chosen: Ensemble Method

Ensemble Method’s Definition

- ML technique combining predictions of smaller/simpler multiple models (e.g. Decision trees, Neural networks) instead of one large/complex model
- Aim is to reduce errors, improve predictive accuracy and reduce overfitting
- Each individual model is either
  - Trained on a different subset of training data

    OR
  - Trained using a different algorithm
- Popular examples of ensemble methods include:
  - Bagging (Boostrap Aggregating) → multiple weak-learners are trained in parallel
  - Boosting → multiple weak-learners are learned sequentially
  - Stacking → trained in parallel, but instead of using simple voting like bagging, another meta-learner is trained on the outputs of weak-learners to learn a mapping from the weak-learners output to the final prediction

Ensemble Method’s Consideration in Design of AI Systems

- ➢ **Diversity of Models** → Individual models differ in underlying algorithms, architectures, and input features, which ultimately reduces the likelihood of multiple models making the same mistakes and hence being less vulnerable to the same attacks
- ➢ **Training Data** → Training data used would be diverse and can more accurately represent different security scenarios and threat landscapes, thereby capturing a wide range of patterns and anomalies which enables the ensembles to detect threats more effectively
- ➢ **Adversarial Robustness** → Cybersecurity systems are susceptible to adversarial attacks, and when designing ensemble methods, adversarial robustness techniques can be utilised to refine the ensemble’s ability to detect and mitigate attacks aimed at exploiting vulnerabilities in the individual models

Ensemble Method’s Consideration in Design of AI Systems

- **Monitoring & Maintenance** → As the threat landscape continues to evolve so does our models, continuous maintenance in the dataset that it is trained on and the complexity of the model is required for the model to detect advance threats to ensure the models long term effectiveness . By actively maintaining and adapting the ensemble techniques, companies can strengthen their cyber defenses ahead of emerging risk
- **Deployment Considerations** → As ensemble methods require more processing power & time to make an inference the model have to be of adequate complexity to detect the advance threats as well as keep up with the mass amounts of data. Thus, factors such as computational resources, scalability and integration with other security systems have to be carefully considered before the deployment of ensemble models

Ensemble Method’s Consideration in Design of AI Systems

- **Evaluation & Validation** → Thorough evaluation and validation of the ensemble's performance are crucial. Comprehensive testing against various attack scenarios, benchmark datasets, and real-world data is necessary to assess the ensemble's effectiveness and robustness. Regular auditing and validation help identify potential weaknesses, biases, or performance degradation.
- **Combination Techniques** → Ensemble methods have various techniques to combine predictions of all the models. Techniques such as majority voting, weighted voting, and as aforementioned, stacking, boosting, and bagging. The choice of combination technique depends on the characteristics of the models, and specific requirements of the cybersecurity system. With the opportunity to evaluate and compare the performance of different combination techniques, the most effective approach for a particular context can be determined.

Ensemble Methods & Its Application to Secure an AI System

- Ensemble methods enhance AI system accuracy by combining predictions from diverse models, reducing errors and minimizing false positives and negatives. The inclusion of various algorithms and architectures in ensembles strengthens security and adversarial robustness, making it more difficult for attackers to exploit vulnerabilities. Regular monitoring, maintenance, and evaluation of ensembles are crucial to ensure effective and secure AI systems. Continuous updates with new threat intelligence and addressing emerging risks bolster resilience against evolving security challenges.
- Overall, ensemble methods provide improved accuracy, security, and adaptability for AI systems through diversity and ongoing vigilance.

Activity 5: Modelling Aforementioned Attack Using Mitre ATLAS Navigator

IDK

-