**Introduction to Adversarial Machine Learning (AML) Risks**

Adversarial Machine Learning (AML) has transformed cybersecurity, emerging as a pivotal element in espionage and warfare. Notable events, such as interference with airport facial recognition technologies and compromises in spam detection systems, highlight the significant dangers associated with AML.

A group of robots dancing

Description automatically generated

Organizations can counteract these threats through strong defensive measures and strategic partnerships, enabling the safe use of AI technologies without an increased susceptibility to AML attacks.

**Understanding Adversarial Machine Learning**

AML operates by exploiting weaknesses in machine learning systems, introducing harmful inputs that lead to malfunctions. Attackers must first dissect the workings of the AI systems before they can alter data to disrupt operations. This is especially concerning in industries like finance, healthcare, and national security, where AI handles critical information.

**Adversarial Machine Learning Threats**

**Data Poisoning:** Altering training data to introduce errors that impair the model's functionality.

**Evasion Attacks:** Creating inputs that AI fails to detect, resulting in unnoticed mistakes.

**Model Stealing:** Duplicating models via reverse engineering to circumvent security protocols.

**Model Inversion:** Extracting sensitive training data from model outputs, risking data privacy.

Adversarial machine learning poses significant threats to AI systems, with real-world examples demonstrating the potential for serious consequences. Here are some key threats and examples:

**Data Poisoning Attacks**

Data poisoning involves maliciously altering the training data to introduce errors and impair the model's performance. For instance, researchers showed that by injecting a few doctored images into a machine learning system for detecting traffic signs, they could trick it into misclassifying stop signs as speed limits.

A diagram of a machine learning model

Description automatically generated

**Evasion Attacks**

Evasion attacks generate adversarial inputs designed to cause AI models to make mistakes while appearing normal to humans. A classic example is adding imperceptible noise patterns to images that cause computer vision models to misclassify them, like mistaking a panda for a gibbon.

Panda Adversarial Example

A collage of images of a dog

Description automatically generated

Adversarial stop signs with minor patterns added could even cause self-driving cars to dangerously misread them.

A stop sign with red and green dots

Description automatically generated

**Model Stealing**

Model stealing or extraction involves probing a deployed AI system to recreate and steal the underlying model, circumventing intellectual property protections. Researchers have demonstrated stealing commercial face recognition and other models in industry-scale black-box attacks.

A diagram of a machine learning process

Description automatically generated

**Model Inversion**

Model inversion aims to extract training data from a model's outputs, violating privacy. For example, researchers showed they could reconstruct recognizable images of people's faces from the output of machine learning models trained on those faces.

These threats underscore the importance of developing robust defenses against adversarial machine learning as AI systems become increasingly critical across many sectors.

**Here are some real-world examples of the major adversarial machine-learning threats:**

**Data Poisoning Attacks**

In 2018, Google revealed that advanced spammers had made large-scale attempts to skew Gmail's spam filtering models by submitting misleading training data to poison the classifiers.

Researchers showed they could poison the training data for Microsoft's anti-malware engine to cause misclassification of malware samples as benign.

**Evasion Attacks**

Researchers added minor patterns to stop signs that caused a self-driving car's computer vision to misclassify them, posing safety risks.

Adversarial eyeglass frames with printed patterns were designed to evade facial recognition systems and cause misidentification.

Cybersecurity firm Skylight Cyber bypassed Cylance's anti-virus by adding strings to malware that tricked the ML model into seeing it as benign.

**Model Stealing**

Researchers stole proprietary models like Google's machine translation system by creating accurate replicas through model extraction techniques.

A proof-of-concept code was demonstrated by stealing the Proofpoint email scoring model through a black-box attack.

**Model Inversion**

Researchers showed they could reconstruct recognizable images of people's faces from the outputs of facial recognition models trained on those faces, violating privacy.

These examples across different application domains highlight the potential real-world impacts of adversarial machine learning threats if defenses are not developed.

**Defensive Strategies Against AML Attacks**

To defend against AML, organizations should implement a multifaceted strategy with human supervision. Data poisoning, evasion attacks, model stealing, and model inversion pose significant threats to machine learning systems. Here are some ways these attacks can be detected and mitigated:

**Detecting Data Poisoning Attacks**

Monitor the training data pipeline for anomalies or outliers that could indicate poisoned data injections.

Use robust data sanitization techniques like RONI (Removing Objects from Datasets) to identify and remove poisoned instances.

Leverage ensemble models trained on disjoint subsets of data to detect disagreements that may signal poisoning.

**Mitigating Evasion Attacks**

Use adversarial training by augmenting the training data with adversarial examples to improve the model's robustness.

Deploy ensemble models that are harder to evade than individual models.

Monitor the model's inputs and outputs to detect adversarial perturbations during inference.

**Preventing Model Stealing**

Use watermarking techniques to verify ownership and detect theft of proprietary models.

Deploy model obfuscation methods that make it harder to extract or reverse-engineer the model.

Limit API access and monitor query patterns to detect attempted model extraction.

**Mitigating Model Inversion Attacks**

Apply differential privacy techniques during training to limit the amount of private information leaked.

Use knowledge distillation to transfer a model to a new architecture, obfuscating the original training data.

Avoid releasing model outputs or confidence scores that may leak too much information.

Developing robust defenses against these threats requires a holistic approach combining techniques across data sanitization, model hardening, monitoring, and privacy-preserving machine learning.

**Impact of AML on Daily Operations**

The repercussions of AML attacks can be severe, costing companies financially, damaging their reputations, and compromising customer data privacy. Risks include:

**Fraudulent Transactions:** Utilizing AML to deceive fraud detection systems, facilitating unauthorized transactions.

**Stock Market Manipulation:** Misleading AI that predicts financial trends, causing losses and market turmoil.

**Misdiagnosis:** Altering outputs from medical imaging AI, resulting in incorrect diagnoses.

**Insurance Fraud:** Convincing AI in insurance processing to approve fraudulent claims.

**Surveillance System Interference:** Disabling facial recognition used in national security.

**Compromising Autonomous Weapons:** Leading autonomous defense systems to target inaccurately.

**Privacy Breaches:** Extracting private information from anonymized data via adversarial methods.

**Information Filter Manipulation**: Altering news algorithm filters to propagate misinformation.

**Autonomous Vehicle Sabotage:** Disturbing the navigation of self-driving cars with manipulated inputs.

**Smart Home Intrusions:** Misleading voice-activated devices in smart homes.

**Legal Evidence Tampering**: Influencing AI systems that process legal documents to affect legal decisions.

**AML Risk Assessment for Enterprises and Government**

Organizations must perform an in-depth AML risk assessment and develop a robust defense strategy. Our structured approach includes:

**AI System Analysis:** Distinguishing between predictive and generative AI to pinpoint vulnerabilities.

**Lifecycle Management:** Securing AI models throughout their development and deployment.

**Understanding Attacker Motives and Capabilities:** Assessing the intentions and abilities of potential adversaries.

**Tailored Risk Assessments:** Carrying out specific risk analyses for various AI applications.

**Strategy Development:** Creating strategies to enhance data security and perform regular security audits.

**Ethical and Legal Considerations:** Considering the ethical and legal ramifications of AML.

Written by Olawale Omoyeni

**References**

Excite Cyber (n.d.). Securing the future: Strategies for adversarial machine learning risk management. Excite Cyber. Retrieved from <https://excitecyber.com/securing-the-future-strategies-for-adversarial-machine-learning-risk-management/>

Khaled, S. (n.d.). The path to resilient AI: Transforming artificial intelligence from fragile to robust. LinkedIn. Retrieved from <https://www.linkedin.com/pulse/path-resilient-ai-transforming-artificial-from-khaled-phd/>

Nevitt, C. (2019, April 4). Why adversarial examples are such a dangerous threat to deep learning. Security Intelligence. Retrieved from <https://securityintelligence.com/articles/why-adversarial-examples-are-such-a-dangerous-threat-to-deep-learning/>

Viso.ai (2021). Adversarial machine learning: Overview, perspectives, and challenges. Viso.ai. Retrieved from <https://viso.ai/deep-learning/adversarial-machine-learning/>

FSISAC (2020). Adversarial AI framework, taxonomy, threat landscape, and control frameworks. FSISAC. Retrieved from <https://www.fsisac.com/hubfs/Knowledge/AI/FSISAC_Adversarial-AI-Framework-TaxonomyThreatLandscapeAndControlFrameworks.pdf>