Cybersecurity Threats and Mitigation Strategies in AI Applications

M Sajjad H. Bhuiyan
Syracuse University

Dr. Joon S. Park
Syracuse University

Learning Objective:

The paper is intended for all audience

- Developers
- Business Managers
- Entrepreneurs
- Anyone interested to understands AI and Cybersecurity

Before we start lets ask a question !!!

Can we trust AI or AI machines?

Popular AI Technologies shaping our future

ChatGPT

Generative AI chatbot

Autonomous Cars

Tesla’s Optimus bot

Fake News

CYBERSECURITY THREATS WITH AI APPLICATIONS

- A) Data Poisoning
- B) Model Theft
- C) Adversarial Attacks
- D) Reverse Engineering
- E) Privacy Leaks

Data Poisoning

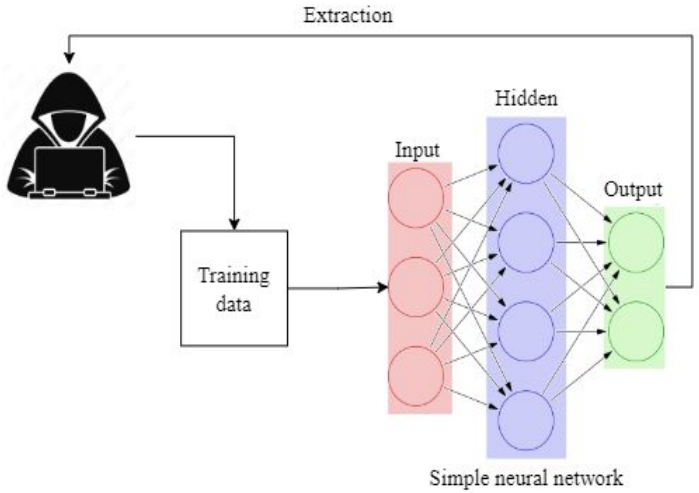


Fig. 1. Data poisoning in action

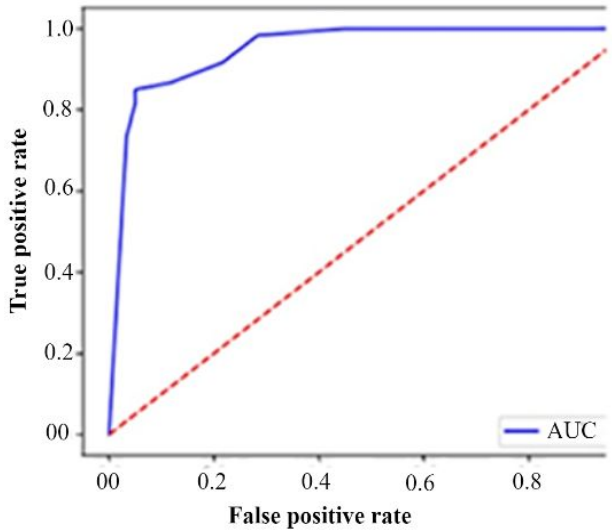


Fig. 2. A Training Data Set with no Data Poisoning

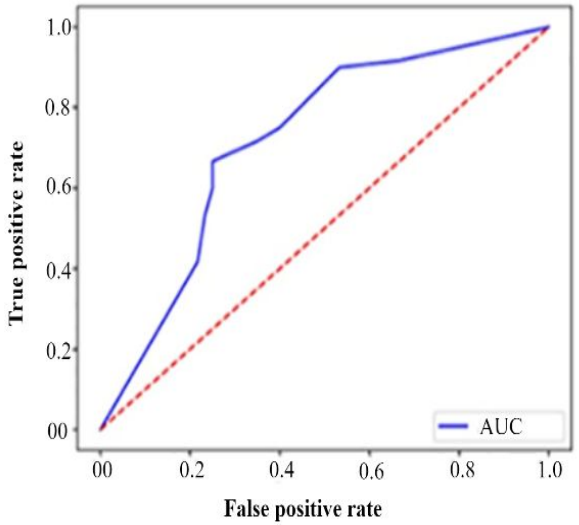


Fig. 3. A training dataset with 25% data poisoning

Model Theft

Unauthorized access to AI models, known as model theft or extraction, allows attackers to replicate and misuse the AI application.

Did you know?

- Back in 2019, Tesla initiated a lawsuit against Guangzhi Cao, a former Autopilot engineer who quit to join Xpeng's autonomous driving team

  In the lawsuit, the automaker claims that Cao downloaded the Autopilot source code to his personal device through Airdrop before leaving and selling it to Xpeng when joining the company.

- AI Model theft in cybersecurity can have severe consequences:

  Compromised Security: compromised model can be used to bypass defenses, results in financial damages and competitive disadvantages

- Addressing model theft requires robust security measures, ongoing monitoring, and breach mitigation strategies. Techniques like model watermarking, differential privacy, and securing multiparty computation can enhance AI model security and reduce theft risks

Adversarial Attacks

Adversarial attacks significantly limit AI in cybersecurity by creating inputs that mislead AI models into making incorrect decisions

Creating Manipulated inputs → AI model make incorrect decision → reduced AI powered solutions

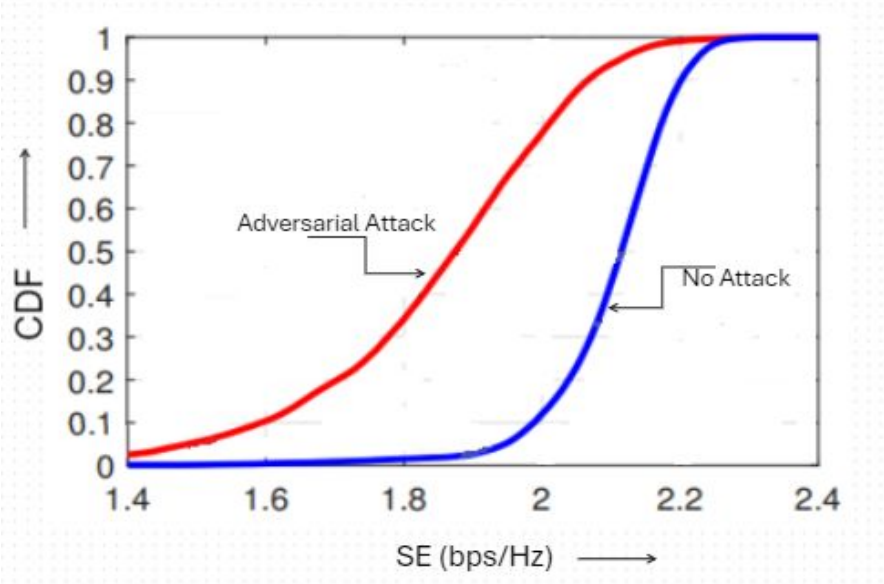


Fig. 5. Effect of adversarial attack on MIMO

Adversarial attacks can have serious consequences:

- Undermining reliability of AI model
- Effectiveness of AI model
- Undetected malicious activities
- Ethical and legal concern

Addressing Adversarial attacks requires adversarial training, robustness checks, and new model architecture

Reverse Engineering

In AI and cybersecurity, reverse engineering involves attackers analyzing AI models to understand their function, identify vulnerabilities, or extract proprietary information.

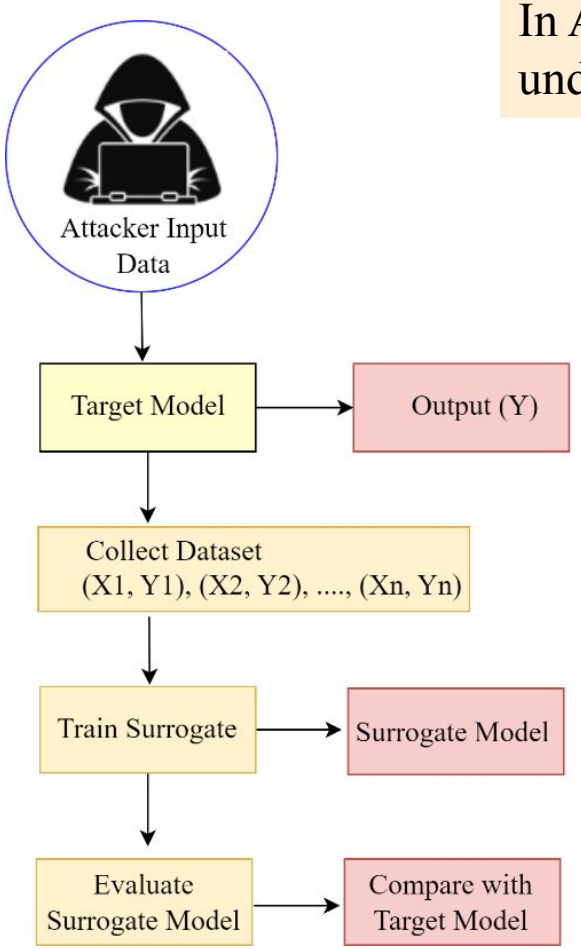


Fig. 6. Model reverse-engineering attack

Consequence of Reverse Engineering

- Expose vulnerabilities
- Attacker can bypass detection or trigger false positive
- Can aid to intellectual property theft
- Attackers can craft inputs to evade detection thus compromise system effectiveness and gain access to sensitive data

Mitigation:

- Multi-layered Approach = technical safeguards + legal protections
- Model obfuscation, Secure enclaves
- Legal measures (e.g., copyright and patents)

Privacy Leaks

What is privacy leaks?

- Unintended disclosures of sensitive or personal information through AI models
- Membership inference attacks might disclose whether specific data was used in training, exposing individual data or past security details.
- Data extraction via prediction APIs can uncover sensitive information about models or their training data, while transfer learning risks privacy by potentially leaking sensitive data in new contexts
- insufficient data anonymization and model overfitting can lead to privacy breaches, exposing personal or sensitive information.

Mitigation:

- Data protection strategies such as minimization, anonymization, and access controls.
- Advanced methods like differential privacy, federated learning, and secure multi-party computation can further protect privacy while leveraging AI.

IMPACT OF GENERATIVE AI IN CYBERSECURITY

Generative AI impact is multifaceted, offering both innovative solutions to enhance security and new challenges that need careful management.

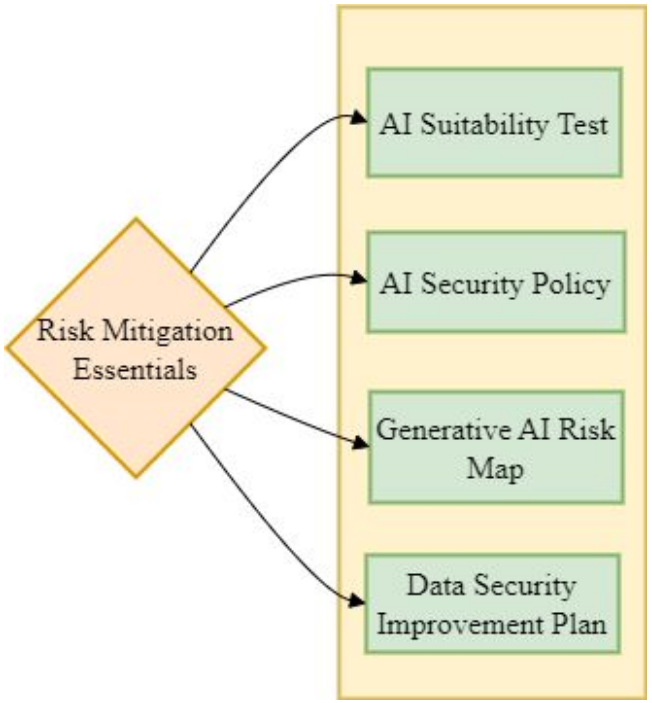
Positive Impacts

- Enhances threat detection by simulating cyber-attacks, identifying vulnerabilities, and strengthening security measures.
- GA automates security tasks by generating configurations and policies, adapting to evolving threats, and reducing manual workload for dynamic security management.
- Generative AI can create realistic phishing simulations to train users to recognize and respond to threats.
- It also supports data privacy by using differential privacy to produce anonymized datasets, protecting individual privacy while maintaining data utility for AI training

Negative Impacts

- Generative AI can produce sophisticated phishing content that is difficult to distinguish from legitimate communications,
- It can also create deepfakes, including realistic images, videos, and audio, leading to impersonation, fraud, and challenges in identity verification.
- Additionally, generative AI can create or modify malware, making it more adaptable and challenging to detect, thus accelerating the cyber arms race.
- If trained on personal data, it may also unintentionally generate outputs with sensitive information and amplify biases in training data, leading to unfair or discriminatory outcomes with privacy and ethical implications.

Generative AI Risk Mitigation



The dual-edged nature of generative AI's impact on cybersecurity and privacy necessitates a balanced approach to its deployment and regulation.

Fig. 7. Generation AI risk mitigation essentials

PRIVACY AND ETHICAL CONCERNS WITH AI APPLICATIONS

A. Privacy, Bias, Security, and Fairness in AI: AI system relies on vast majority of data. Transparent data collection processes and precise consent mechanisms are essential to ensure the ethical handling of personal information. Biases in training data can result in discriminatory outcomes

Mitigation: prioritize bias mitigation, using fairness-aware machine learning techniques and diverse datasets. Proper accountability frameworks and compliance with regulations such as GDPR (General Data Protection Regulation)

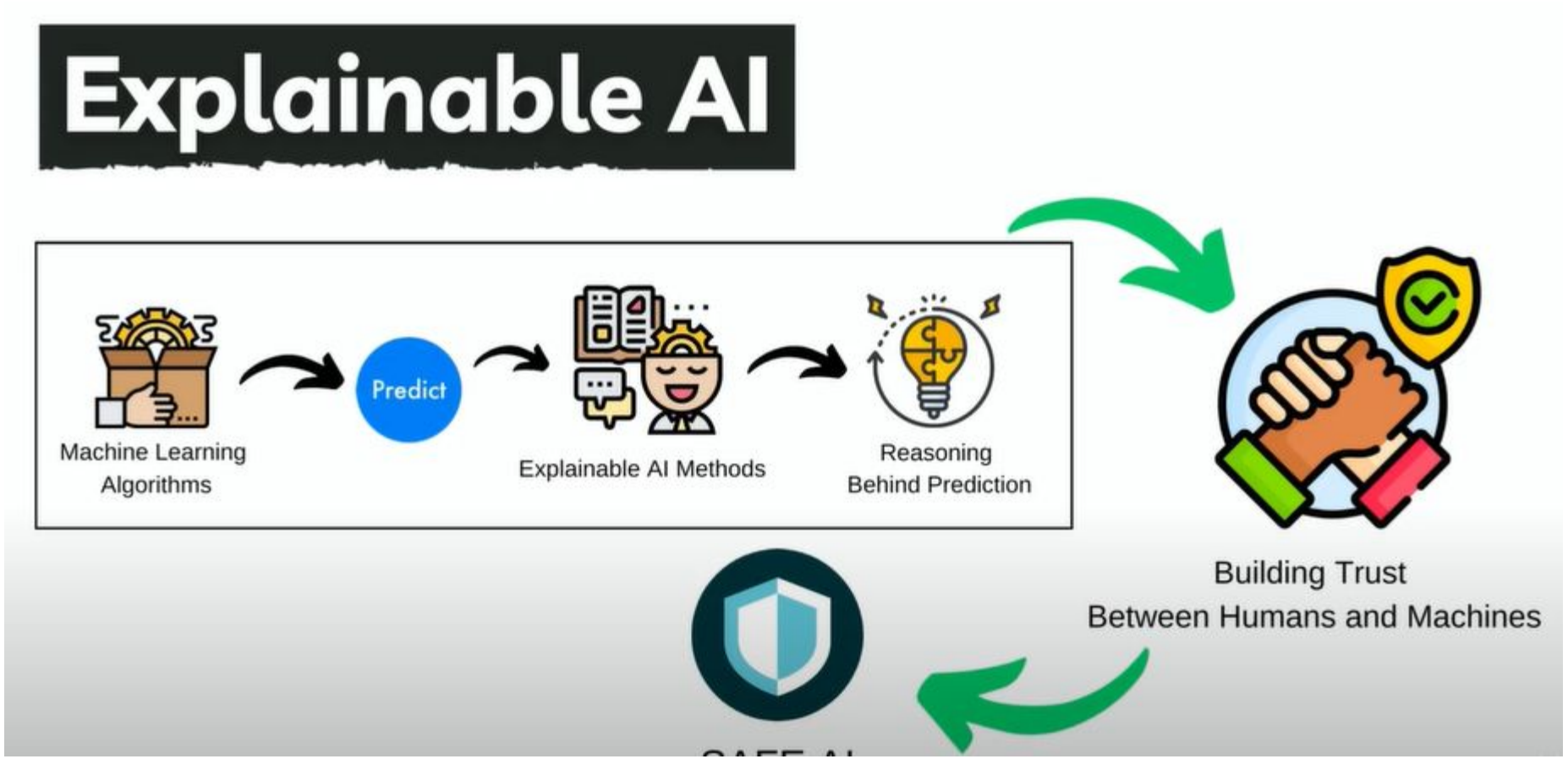
B. Transparency and Human Control in AI: AI's opaque decision-making processes challenge transparency. As AI takes on more decision-making roles, there are growing ethical concerns about eroding human control. AI should assist, not replace, human decision-making, ensuring that responsibility remains with humans rather than machines.

Mitigation: (XAI) systems that non-experts can understand and assess. Algorithmic transparency is crucial for ensuring fairness, allowing users to trace decision-making methods and data sources.

C. Security and Ethical AI Development: The rapid adoption of AI introduces new security risks, making robust security measures vital to prevent vulnerabilities and misuse. AI in surveillance by governments or corporations raises ethical concerns about privacy and individual freedoms.

Mitigation: Access controls, regular security assessments, and incident response planning. Ethical AI development also requires educating developers on cybersecurity, privacy, and ethical risks during system design, ensuring compliance with regulations and ethical principles.

VI. ENHANCING AI SECURITY WITH EXPLAINABLE AI



VI. ENHANCING AI SECURITY WITH EXPLAINABLE AI

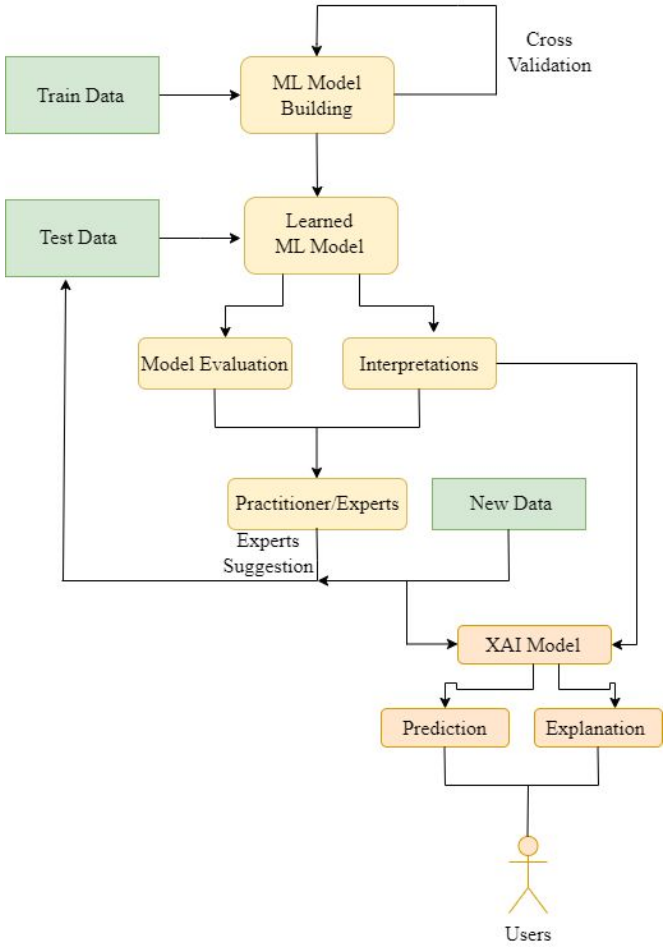


Fig. 8. Framework of XAI model

Explainable AI promote transparency, allows greater interpretability of AI models and insights into decision-making processes.

XAI techniques clarify the decision-making processes of ML models, providing transparency that is critical for cybersecurity applications where understanding alerts are crucial for trust and accurate decision-making.

XAI strengthens cybersecurity by improving attack detection and promoting effective communication between AI systems and human analysts in AI models.

VII. THE FUTURE OF AI CYBERSECURITY: EMERGING TRENDS

The future of AI cybersecurity will be shaped by advancements in AI technologies, innovative solutions, and collaboration among stakeholders to address emerging threats as follows.

A. AI-Powered Cyber Defense

B. Adversarial Machine Learning

C. Explainable AI in Cybersecurity

D. Privacy-Preserving AI Security

E. Cybersecurity for AI Applications

F. Regulatory and Ethical Frameworks

Overall, the future of AI cybersecurity will be characterized by constant innovation, adaptation, and collaboration to address evolving cyber threats, safeguarding digital assets, infrastructure, and individuals in an increasingly AI-driven world.

VIII. CONCLUSION

Cybersecurity concerns in AI applications are complex and multifaceted, encompassing technical, ethical, and regulatory challenges.

Protecting AI systems from threats demands a comprehensive strategy, including

Strong data protection

Resilience against adversarial attacks, and

Careful consideration of the ethical dimensions of AI security.

To prevent AI exploitation, robust security measures, continuous monitoring, and adversarial training are essential in addressing vulnerabilities.

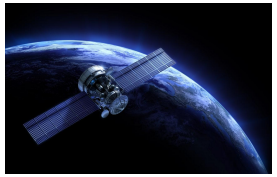
Explainable AI can play a key role by offering transparent and interpretable insights into model decisions, fostering trust and accountability.

Collaboration among researchers, industry experts, and policymakers is crucial to developing standards, guidelines, and best practices for securing AI applications and ensuring their safe, beneficial use in society.

I am going to end my presentation with IROBOT movie last few seconds

Movie: IROBOT 2004

Finally Can we Trust AI?



TESLA

Too much Power!!!!!

With uncontrolled AI our future can be at risk

Dictators/Criminals can have too much power

Carelessness/Mistakes can create undesired outcomes

Thus, we must pay undivided attention to

Cybersecurity Threats and Mitigation Strategies in AI Applications

Thank you!