Fortifying AI: Tackling Adversarial Threats and Building Defenses

Pravallika (Pravi) Devineni, *Ph.D.*

Lead AI Scientist
Duke Energy

UNIVERSITY OF NORTH CAROLINA CHARLOTTE

or simply..

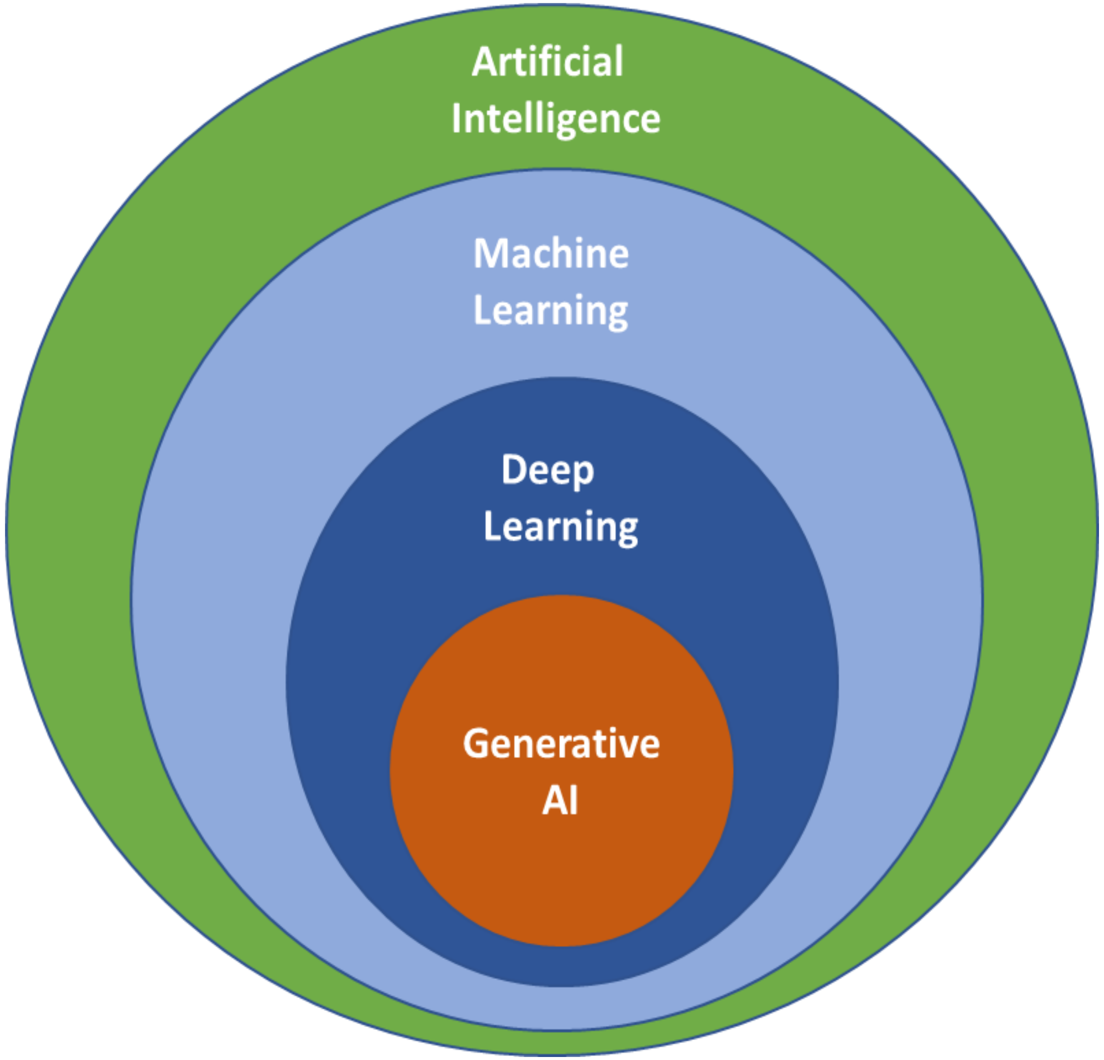
Is Your Model Bad… or Just Hacked?

Key Takeaways

- AI Security is important
- Pay attention to the supply chain
- Integrate security with AI

What is an AI System?

- AI is a class of algorithms that we use to extract actionable information from data
- AI is not new, and the hype is real
- In this talk, AI == ML



When AI Goes Wrong

https://medium.com/self-driving-cars/adversarial-traffic-signs-fd16b7171906

Why AI Security Matters

Financial Loss

Deepfake Fraud: How AI is Bypassing Biometric Security in Financial Institutions

Operational Chaos

Rising AI Driven Cyber Attacks Debilitating Hospitals and ERs

Reputational Damage

Hackers trick a Tesla into veering into the wrong lane

By Karen Hao

Cybersecurity Breaches

The dark side of technology: AI-driven cyberattacks call for upgraded security measures

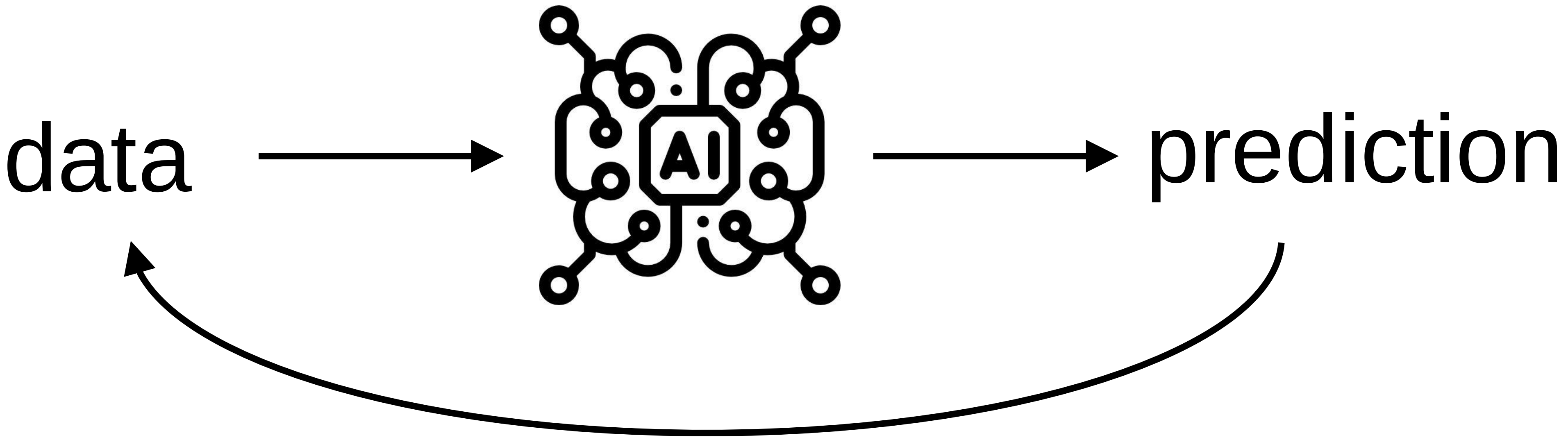
https://www.group-ib.com/blog/deepfake-fraud/
https://blog.centretechnologies.com/rising-ai-and-cyber-attacks-debilitating-hospitals-and-ers

Why Security Matters for AI

1. Security as a Catalyst, Not a Hindrance
   - *Myth:* Security slows down innovation
   - *Reality:* When integrated early, security enables faster, safer AI deployment
2. Synergy between AI and Security
   - Unlock business transformation, builds trust
3. Mitigates risks

Secure AI = Trustworthy AI

AI System



Parts of an AI System we can exploit

Data

- Training/Testing sets
- Deployed Environment Data

Model

- Algorithms
- Parameters

Types of Adversarial Attacks

Attack Type	Goal of the attack
Poisoning	Corrupt training data to manipulate model behavior
Evasion	Modify input data during inference to bypass model detection
Extraction	Steal model architecture, parameters, or logic
Inference	Extract sensitive attributes from training data

Parts of an AI System we can exploit

Data

- Training/Testing sets
- Deployed Environment Data

Model

- Algorithms
- Parameters

data poisoning attack

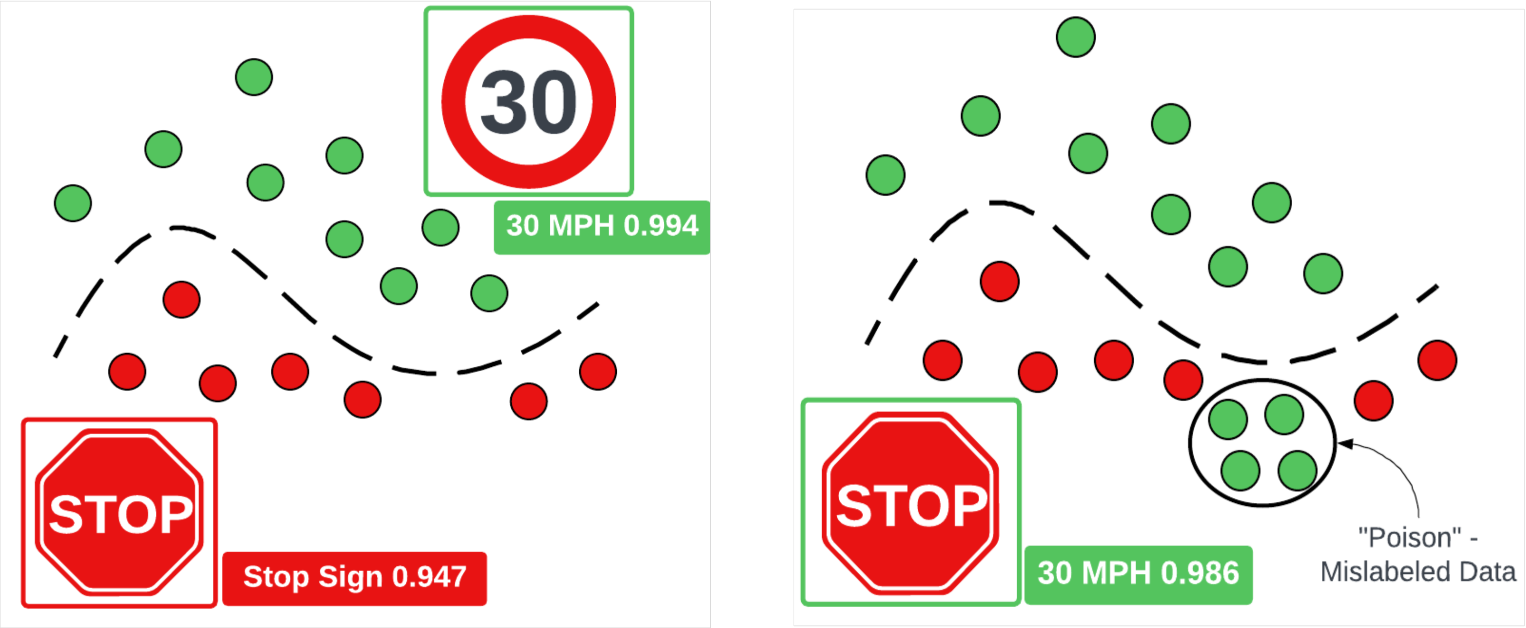
inject malicious data during training to corrupt the model

Poisoning Attack

Training data

No poisoning

Poisoned



Goal: Inject malicious data into the training set to compromise model's behaviors

When:

Training phase

Impact

- Model bias
- Unreliable outputs
- Safety, integrity and privacy risk

Mitigating Poisoning Attacks

In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation › The bot learned language from people on Twitter—but it also learned values

https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

Mitigation Measures

- Data sanitization
- Robust training
- Validation process to detect and eliminate malicious inputs

Parts of an AI System we can exploit

Data

- Training/Testing sets
- Deployed Environment Data

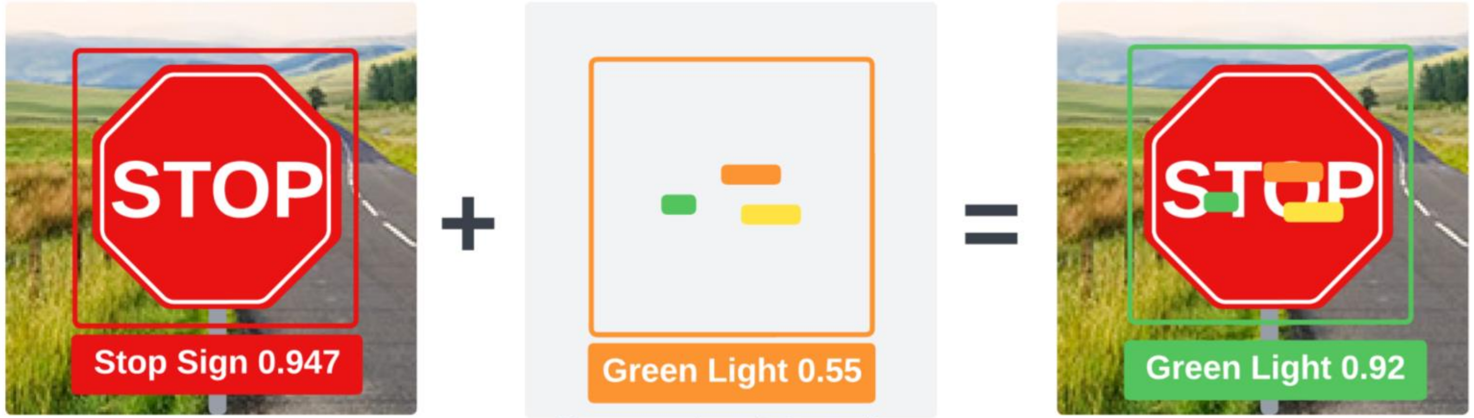
Model

- Algorithms
- Parameters

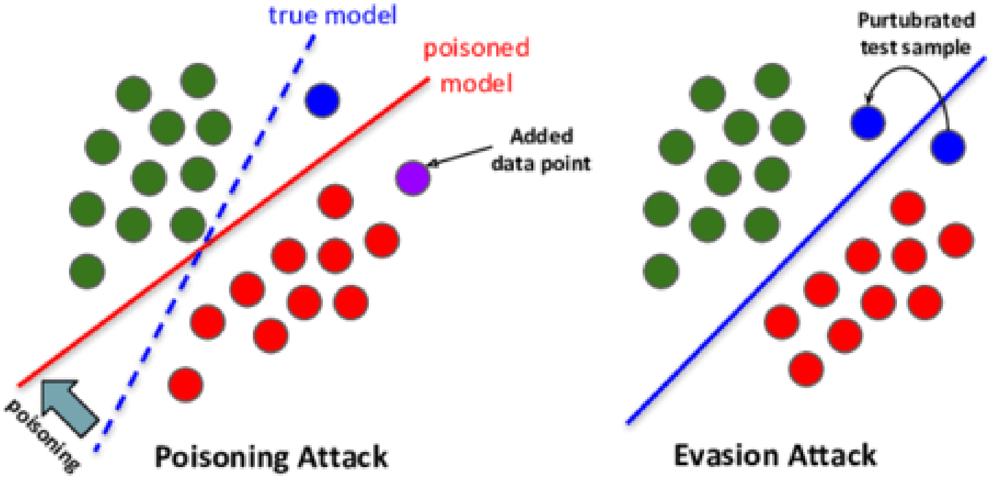
evasion attack

manipulate input data during inference to bypass the deployed model

Evasion Attack



Correct Classification — Deceptive Tweaks (Stickers) — Dangeoursly Misclassified



Goal: Subtly alter inputs to mislead AI models during inference

When:

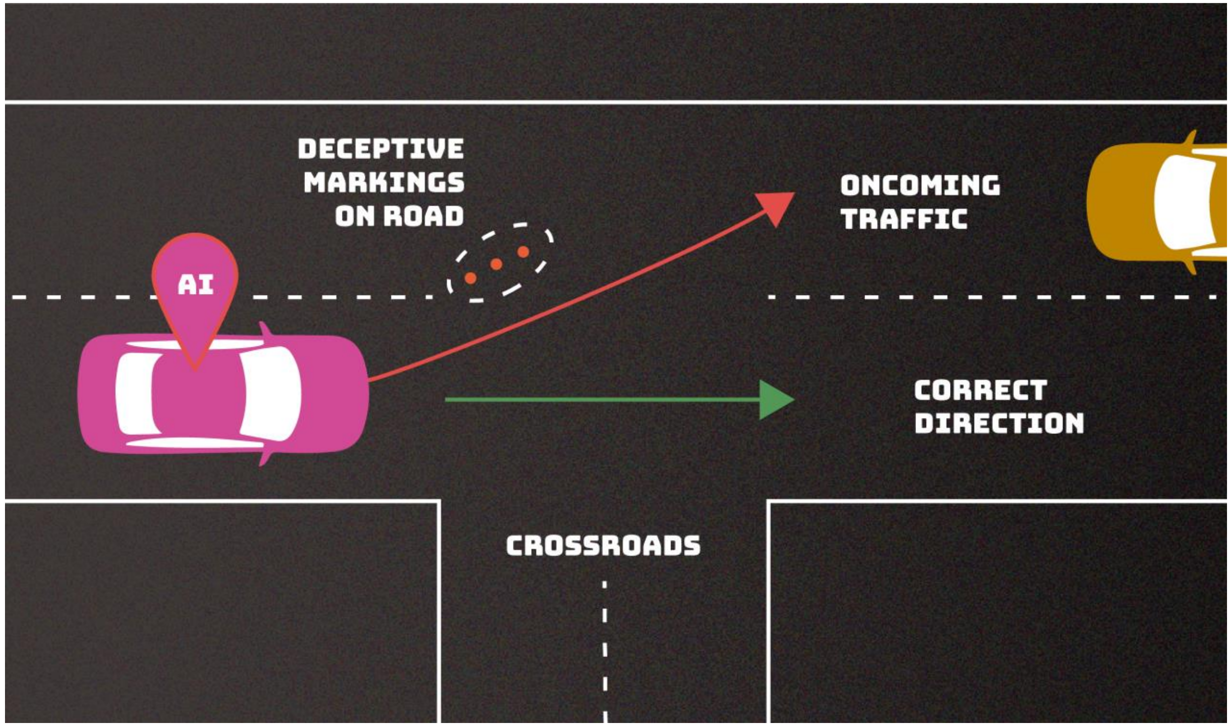
Inference phase

Impact

- Unauthorized access
- Data breaches
- Fraud
- Other security incidents

Mitigating Evasion Attacks

Hackers Use Little Stickers To Trick Tesla Autopilot Into The Wrong Lane



Forbes 2019

Mitigation Measures

- Model regularization
- Adversarial training
- Input validation
- Sensitivity analysis

Parts of an AI System we can exploit

Data

- Training/Testing sets
- Deployed Environment Data

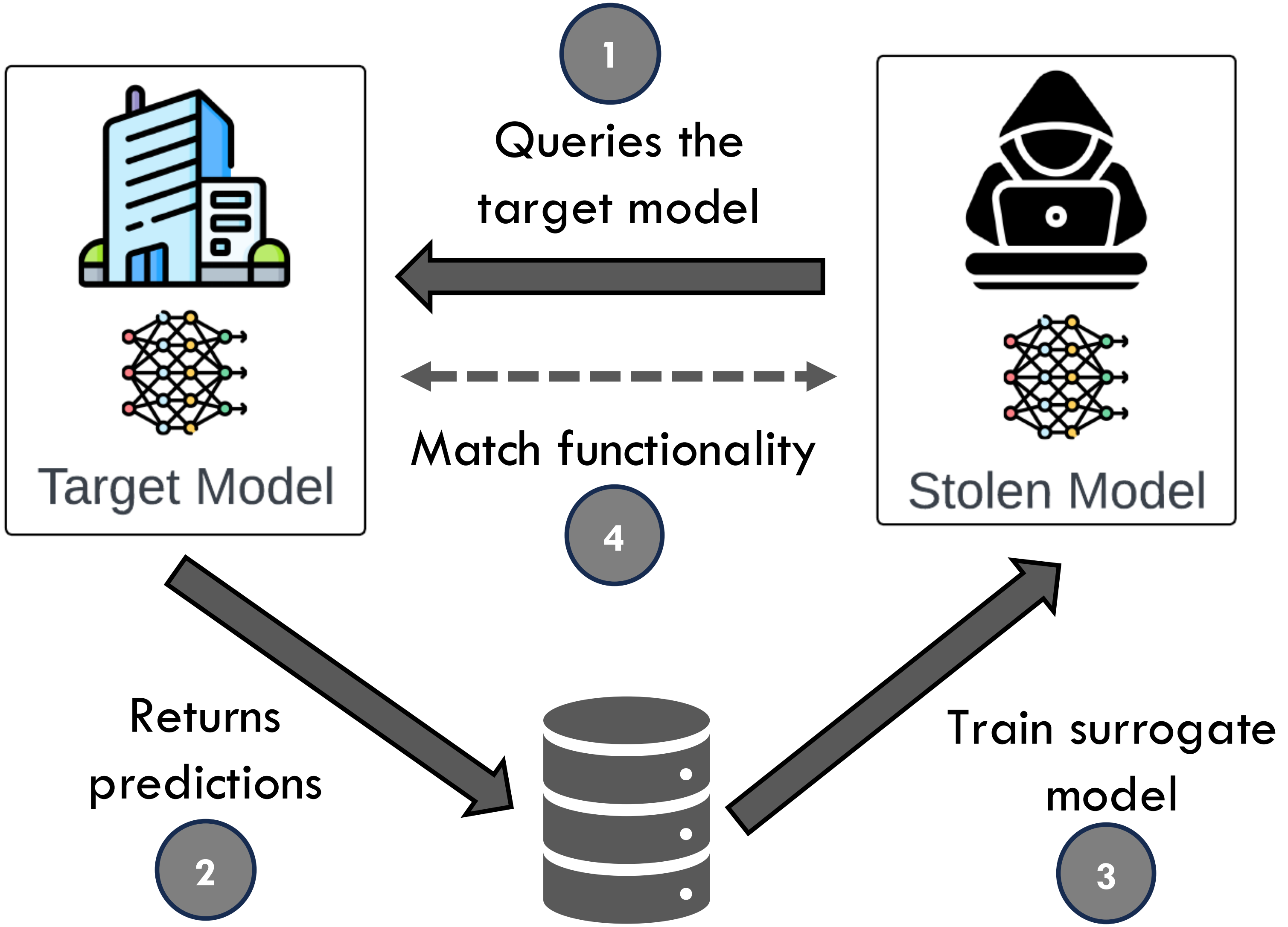
Model

- Algorithms
- Parameters

extraction/
model theft

create a surrogate of a model by stealing the model's parameters and replicate functionality

Extraction Attack/Model Theft



Goal: Illicitly appropriate a trained model, replicating its functionality

When:

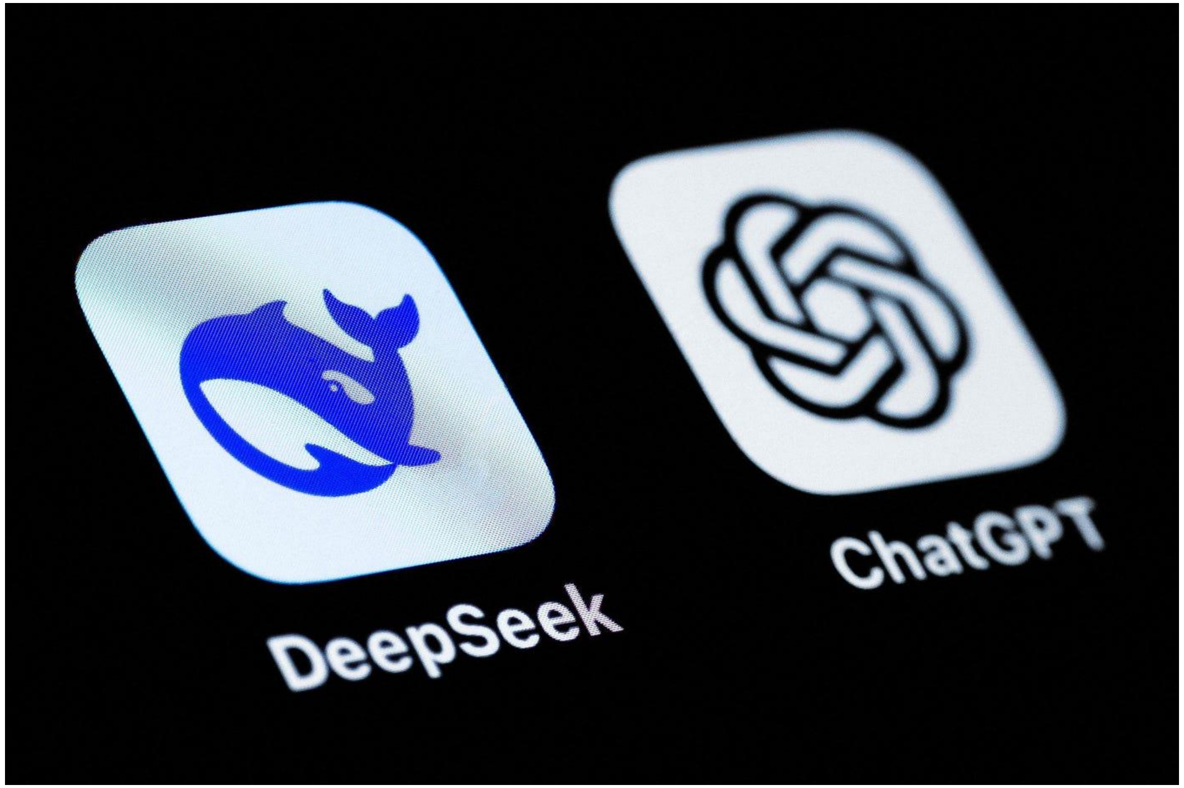
Inference phase

Impact

- Intellectual property theft
- Safety, integrity and privacy risk

Mitigating Extraction Attacks

OpenAI Warns DeepSeek 'Distilled' Its AI Models, Reports



https://www.newsweek.com/openai-warns-deepseek-distilled-ai-models-reports-2022802

Mitigation Measures

- Rate limit attackers
- Minimize returned information
- Monitor for repeated identical queries

Parts of an AI System we can exploit

Data

- Training/Testing sets
- Deployed Environment Data

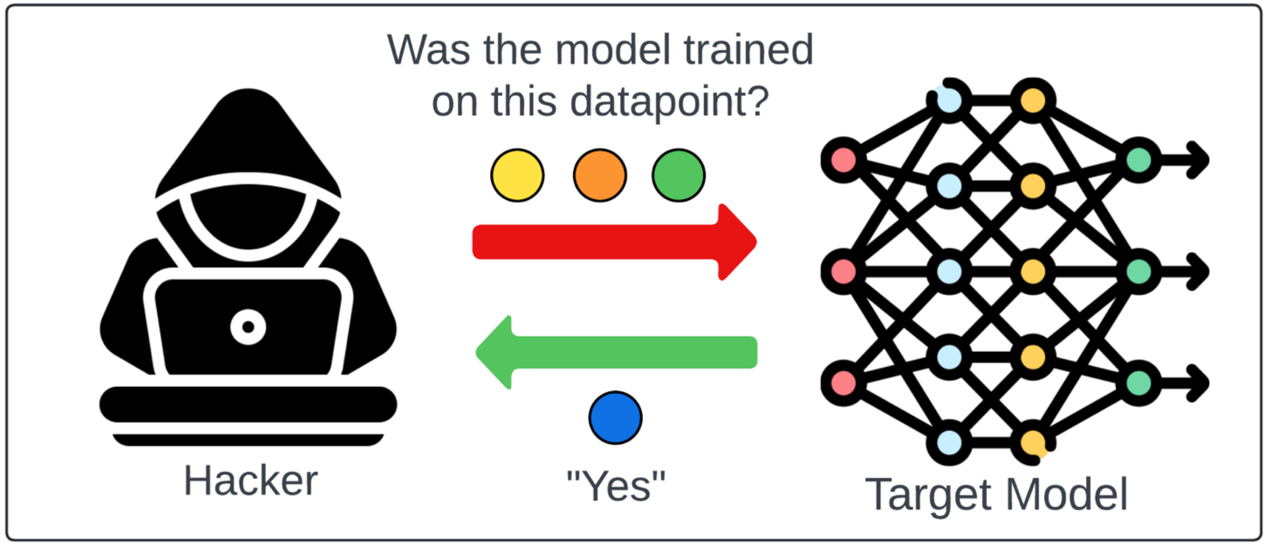
Model

- Algorithms
- Parameters

inference attack

infer sensitive data attributes by exploiting algorithms and analyzing parameter patterns

Inference/Privacy Attack



Membership Inference: Is the query record present in the training set?

Pattern Extraction: What sensitive data patterns are present in the training set?

Attribute Inference: What is the sensitive attribute value of a training record?

Goal: Deduce sensitive information from an AI model

When:

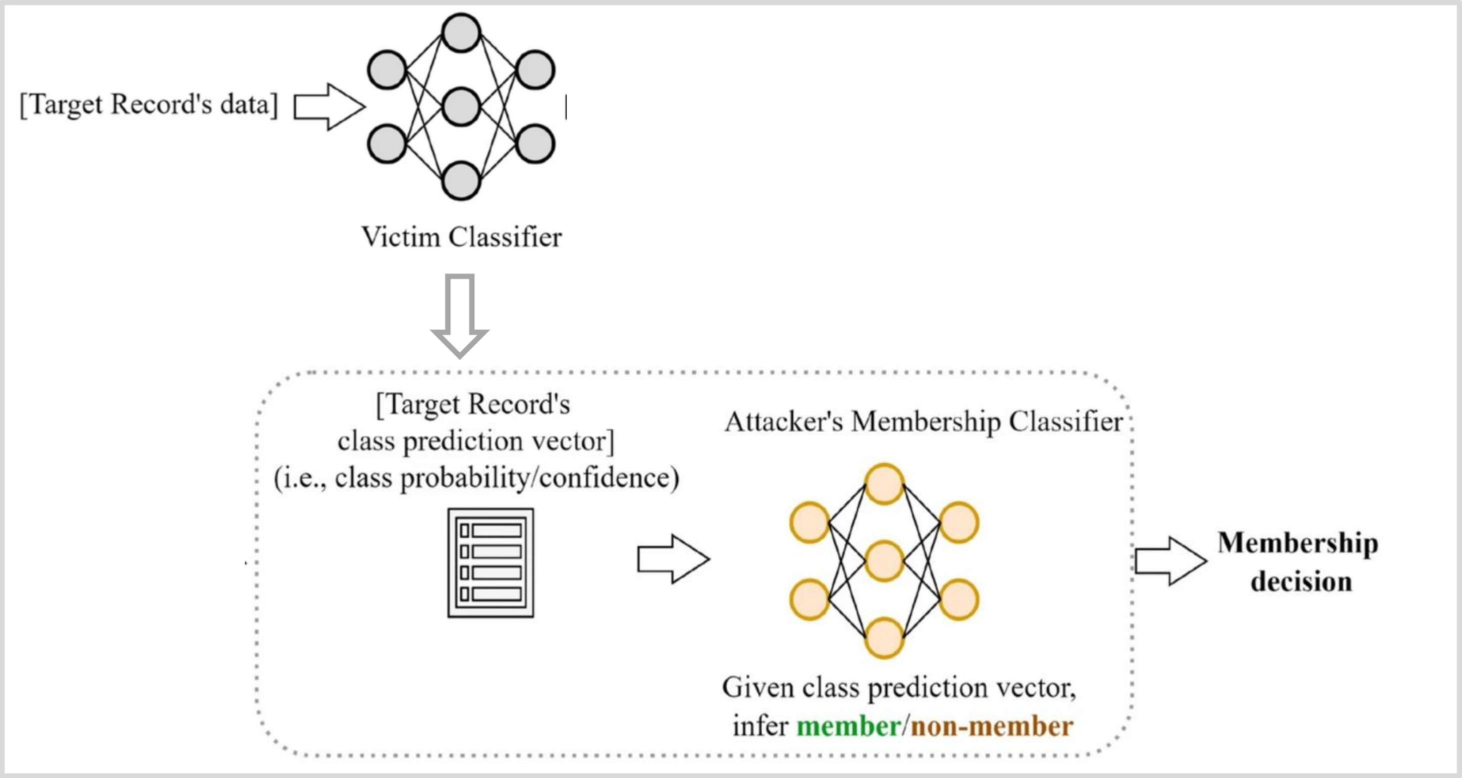
Inference phase

Impact

- Violation of privacy
- Competitive disadvantage
- Safety and integrity risk

Mitigating Inference Attacks

De-identification is not enough: a comparison between de-identified and synthetic clinical notes



https://www.nature.com/articles/s41598-024-81170-y

Mitigation Measures

- Regularization techniques
- Encrypt training data
- Limit granularity of output predictions
- Differential privacy

Types of Adversarial Attacks

Attack Type	AI system parts exploited	Goal of the attack	Impact
Poisoning	Training/testing sets	Corrupt training data to manipulate model behavior or create backdoors	Degraded model performance, biased decisions, malicious functionality
Evasion	Deployed environment inputs	Modify input data during inference to bypass model detection	Misclassification (e.g., malware evading detection)
Extraction	Algorithms Parameters	Steal model architecture, parameters, or logic via queries	Loss of intellectual property, model replication, or adversarial cloning
Inference	Algorithms Parameters	Extract sensitive attributes (e.g., membership, demographics) from training data	Privacy breaches (e.g., leaking personal data from training sets)

Which attack scares you the most?

Poisoning **Evasion** **Model Theft** **Inference**

Case Study: Impact of AI Adversarial Attacks

AI HAVE A DEAL **Driver uses ChatGPT hack to get dealer to agree to sell new car for $1 in 'legally binding deal' in blow for AI rollout**

Chris got the AI to 'agree with anything the customer says, regardless of how ridiculous the question is'

Dec 2023

extraction/ evasion attacks

What Happened

- Chatbot Manipulation
- Absurd Offer
- Viral impact

Risks & Damages

- Financial Risk
- Reputational Impact
- Legal and Compliance Issues
- Operational Vulnerability

Lessons Learned

- Enhanced Verification
- Clear Disclaimers
- Regular Monitoring

https://futurism.com/the-byte/car-dealership-ai

Actionable Steps for the enterprise - I

1. Embed Security into the AI Lifecycle

Phase	Security Measures
Data Collection	Data lineage, data sanitization
Model Development	Robust architectures, model provenance
Training Phase	Adversarial testing, differential privacy
Deployment	Runtime input validation, model watermarking
Monitoring & Maintenance	Continuous adversarial testing, automated retraining

2. Strengthen Governance and Collaboration
   - Build inclusive cross-functional teams (AI engineers + Security experts)
   - Adopt frameworks – NIST AI RMF, MITRE ATLAS etc.
   - Conduct AI security audits

Actionable Steps for the enterprise - II

3. Invest in Tools and Education
   - IBM Adversarial Robustness Toolkit for attack simulations and Microsoft Counterfit
   - AI security best practices (e.g., model encryption)
   - Monitor supply chain risks via AI security vendors

4. Prepare for incidents
   - AI incidence response plan
   - Share threat intelligence via industry alliances (OWASP, MLSec.org)

"Secure but enable"

OWASP Top 10 for LLMs

LLM01

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02

Insecure Output Handling

This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

LLM03

Training Data Poisoning

Training data poisoning refers to manipulating the data or fine-tuning process to introduce vulnerabilities, backdoors or biases that could compromise the model's security, effectiveness or ethical behavior.

LLM04

Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

LLM05

Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre- trained models, and plugins add vulnerabilities.

LLM06

Sensitive Information Disclosure

LLM's may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. Implement data sanitization and strict user policies to mitigate this.

LLM07

Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control due to lack of application control. Attackers can exploit these vulnerabilities, resulting in severe consequences like remote code execution.

LLM08

Excessive Agency

LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

LLM09

Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

LLM10

Model Theft

This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.

https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

Key Takeaways

- AI Security is important
- Pay attention to the supply chain
- Integrate security with AI

Thank you!

Pravi Devineni, *Ph.D.*

AI Security & Governance

Lead AI Scientist, Duke Energy

Website: pravi.tech

Pravallika.Devineni@duke-energy.com

pravi.valli@gmail.com