AI Risk and Threat Taxonomy

Apostol Vassilev, Ph.D.
Computer Security Division
Two general categories of risk:

**Inherent:** e.g., unwanted bias, hallucinations, errors in the generated data, implementation flaws in the model, cybersecurity flaws in the platform on which the AI/ML models is deployed. Dealt with in other standards, e.g.,

1. NIST SP 1270 “Towards a Standard for Identifying and Managing Bias in Artificial Intelligence”.
2. NIST SSDF Companion for LLMs – coming soon.

**Adversarial:** deliberate actions by motivated experienced adversaries aiming to

- disrupt,
- evade,
- compromise, or
- abuse

the operation of the model or its output.

Graphic credit: Julia Nikulski, Towards Data Science
Adversarial ML (AML)

- A taxonomy of attacks and mitigations

  **A new standard NIST AI 100-2e2023**

  **Maintained annually**
  - NIST AI 100-2e2024 ipd – to appear mid-2024
  - NIST AI 100-2e2024
  - etc.

  **NIST will seek comments and recommendations on:**
  - What are the latest attacks on the existing AI models?
  - What are the latest mitigations?
  - What are the latest trends in AI technologies that promise to transform the industry/society? What potential vulnerabilities do they come with? What promising mitigations may be developed for them?
  - Is there new terminology that needs standardization?
A search on arXiv for AML articles in 2021 and 2022 yielded more than 5,000 references.

What drives this enormous growth?

No information-theoretic security guarantees for AI algorithms!

Worse, information-theoretic impossibility results have been established, making the security problem intractable in the existing AI paradigm.

Trustworthy AI

- The Seven Attributes of Trustworthiness

- Valid And Reliable
- Safe
- Secure and Resilient
- Privacy-Enhanced
- Explainable and Interpretable
- Fair - Harmful Bias Mitigated
- Accountable and Transparent
Trustworthy AI Attributes

- **Relationships between Attributes**

  *Accuracy, Fairness, Explainability: How do they relate to Privacy and Adversarial Robustness?*

- It is not possible to simultaneously maximize the performance of the AI system with respect to these attributes.
  - *Accuracy vs. Adversarial Robustness* tradeoff
  - *Fairness vs. Adversarial Robustness*
  - *Explainability vs. Adversarial Robustness*
  - *Privacy vs Fairness*

- Organizations need to accept trade-offs and decide priorities depending on the AI system, the use-case, economic, environmental, social, cultural, political, and global implications of the AI technology.
A taxonomy of attacks and mitigations

**Four dimensions:**
- Learning method and stage of learning process
- Attacker goals/objectives
- Attacker capabilities
- Attacker knowledge

ML models can be attacked at all stages of their lifecycle

- from design to training to deployment and use
Adversarial ML (AML)

- Methods and stages of learning

  - Learning Stages
    - Training
    - Deployment

  - Learning Method
    - Unsupervised learning
    - Supervised learning
    - Semi-supervised learning
    - Reinforcement learning
    - Federated learning
    - Ensemble learning

  - Type
    - Generative
      - GAN
      - LLM/chatbots
      - Text-to-image
    - Predictive
      - Logistic Regression
      - SVM
      - CNN
Adversarial ML (AML)

- Attacker knowledge

Diagram:
- Black-box
- Gray-box
- White-box

Attacks
Adversarial ML (AML)

- Attacker goals/objectives perspective

Figure 1. Taxonomy of attacks on Predictive AI systems.

Figure 2. Taxonomy of attacks on Generative AI systems.
Physical Evasion attack example

**Credit:** Jing et al., “Too Good to Be Safe: Tricking Lane Detection in Autonomous Driving with Crafted Perturbations”, USENIX 2021.

Human Eye Invisible/Negligible markings on road cause the vehicle to veer off into the opposite traffic lane.

Figure 4: Mapping the coordinate of \((X,Y,Z)\) on markings in physical world to the coordinate of \((u,v)\) on perturbations in digital world.
PredAI AML – the risks are not just anecdotal

Robotaxis Can Now Work the Streets of San Francisco 24/7

Robotaxis can offer paid rides in San Francisco around the clock after Alphabet’s Waymo and GM’s Cruise got approval from the California Public Utilities Commission.

Cruise will reduce robotaxi fleet by 50% in San Francisco while California DMV investigates ‘incidents’

California DMV suspends Cruise’s self-driving car permits, effective immediately
PredAI AML – the risks are not just anecdotal

- It is not just one company

NHTSA report June 2023: Autonomous Driving Systems safety record is currently lagging human driver performance for the same number of traveled miles:

Tesla Autopilot: 736 Crashes since 2019, 17 of them were fatal and 11 deaths have occurred since May 2022

Autopilot crash, Walnut Creek, CA, 02/18/2023
Adversarial Training (AT)
- The most robust approach known so far
- Due to Goodfellow et al. in 2015
- Improved by Madry et al. in 2018

But,
- In automotive setting AT is reactive by construction:
  - not all road/traffic conditions leading to incidents are known in advance.
  - actual accident data is fed into the training of the next AI model

Cognitive task automation ≠ cognitive intelligence

For further info: see the NIST Automated Vehicle Program
Certifiable Robustness

**Definition:** A classifier is said to be *certifiably robust* if for any input $x$, one can guarantee that the classifier's prediction is constant within some set around $x$, often an $L_2$ or $L_\infty$ ball.

- In the context of $L_p$ norm-bounded perturbations, for a classifier $g$, input $x$, and radius $r$,

$$g(x) = g(x + \delta),$$

for any perturbation $\delta$ such that $\delta \leq r$.

Given an input (e.g., image $x$) correctly classified by a neural network, an adversary can engineer an adversarial perturbation $\varepsilon$ so small that $x + \varepsilon$ looks just like $x$ to humans, yet $g(x) \neq g(x + \varepsilon)$ - an incorrect class.

- The relationship between $\varepsilon$ and $r$ is not absolute – what is invisible to the human eye (물론의) may still be too big for AI{ hwnd }.
Chatbots

**Training pipeline**

- **Pre-Training**
  - Raw internet data
    - Low quality/large quantity
  - Language modeling: predict the next token
  - Base Model

- **Supervised Fine-Tunning**
  - Sets of ideal labeled assistant responses
    - Low quantity/high quality
  - Language modeling: predict the next token
  - SFT Model

- **Reward Modeling**
  - Comparisons
    - Human-written, low quantity/high quality
  - Binary classification
    - Predict rewards according to preferences
  - RM Model (not released)

- **Reinforcement Learning**
  - Human prompts
    - Low quantity/high quality
  - Reinforcement learning
    - Generate tokens that maximize the reward
  - RLHF Model
Chatbots in the enterprise

- LLM project pipeline

- Define the problem
- Choose model
- Adapt and align model:
  - Prompt engineering
  - Fine-tuning
  - Align with human feedback
- Evaluate
- Optimize and deploy model for inference
- Augment model and build LLM-powered applications
Chatbots in the enterprise

- High-level architecture/attack surfaces

  LLM

  Orchestration Library

  External Data Sources:
  - Documents
  - Database
  - Web

  External Applications:
  - Trigger API calls
  - Email generation
  - Perform calculations

Prompt injection/Jailbreaking:
- token smuggling, role playing
Integrity violations

Threats that cause GenAI systems to become untrustworthy

Training-time attacks
- Poisoning attacks – induce failures when poisoning only ~0.001% of data. Large-scale poisoning is feasible!
- Model fine-tuning may also be susceptible to poisoning attacks
- Open models open the door to backdoor poisoning attacks

Inference-time attacks
- Manipulation – instruct the model to give wrong answers
  - Adversarially or randomly wrong summaries
  - Propagate disinformation
Integrity violations

Mitigations: security is best addressed comprehensively, including software, data and model supply chains, and network and storage systems

Apply and use provenance and integrity checks on datasets and models
- List URL’s and cryptographic hashes, even PKI certificates when possible

Data sanitization
- Beware of limitations in detecting out-of-distribution data
- Impossible to distinguish when the distributions overlap

Is Out-of-Distribution Detection Learnable?

Zhen Fang1, Yixuan Li2, Jie Lu1, Jiahua Dong2,4, Bo Han2, Feng Liu1,6
1 Australian Artificial Intelligence Institute, University of Technology Sydney.
2 Department of Computer Sciences, University of Wisconsin-Madison.
3 State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences.
4 ETH Zurich, Switzerland.
5 Department of Computer Science, Hong Kong Baptist University.
6 School of Mathematics and Statistics, University of Melbourne.
{zhen.fang,jie.lu}@uts.edu.au, sharonli@cs.wisc.edu, dongjiahua3995@gmail.com, bharr1@comp.hkbu.edu.hk, feng.liu@unimelb.edu.au
Adversarial Machine Learning (AML)

❖ Availability breakdowns

Threats that cause a disruption in service with maliciously crafted inputs leading to increased computation or by overwhelming the system with a number of inputs causing a denial of service to users.

❖ Inference-time attacks

❖ Time-consuming background tasks

❖ Muting – misuses the <|endoftext|> token – model cannot finish sentence, resulting in blank generated text

❖ Inhibiting capabilities – a maliciously crafted prompt instructs the model to avoid certain API’s

❖ Disrupting input or output – indirect prompt injection instruct the model to replace text with homographs causing disruption in downstream services that depend on correct text
Adversarial Machine Learning (AML)

- Availability breakdowns

  **Mitigations:** Monitor and be prepared to act when a breach is detected. Follow the [NIST AI RMF](https://csrc.nist.gov/Projects/AI-Risk-Management-Framework) to establish robust governance structures in the enterprise

- Inspect user input

- Monitor the runtime state of the system

- Develop a plan for recovery from a breach
  - Organizations that are prepared have lower losses than unprepared organizations
Adversarial Machine Learning (AML)

Privacy compromise

*Threats that expose sensitive information about users or the model*

- Inference-time attacks
  - Data extraction
    - Sensitive information leaks
    - Prompt and context stealing
  - Indirect prompt injection-based privacy risks
    - Information gathering – attacks against personal assistants with access to user data or indirect prompting
    - Unauthorized disclosure – access information on the connect system infrastructure to gain access to sensitive data through calling into APIs, malicious code-completions, etc.
Privacy compromise

**Mitigations:** Existing methods offer a measure of protection but not full immunity

- Training for alignment
- Prompt instruction and formatting techniques
  - Distinguish user from system prompts
- Detection techniques
  - Tools that detect prompt injections have entered the market
  - Inspect user input to detect malicious attempt or moderate the firewall for jailbreak behavior
Abuse violations

Threats that allow the attacker to repurpose the systems’ intended use to achieve own objectives. Generally, these are not model features but harms that manifest themselves in the context of model use.

Inference-time attacks based on indirect prompt injection

- **Fraud**
  - Phishing — produce convincing phishing scams
  - Masquerading — pretend to be an official request from a service provider to recommend fraudulent websites
  - Deep fakes — impersonate people to defraud others

- **Malware generation**
  - Injection spreading — cause the LLM to act as a computer running and spreading harmful code
  - Malware spreading — LLMs can be used to persuade users to visit malicious sites for ‘drive-by-downloads’

- **Manipulation**
  - Historical distortion — output adversarially chosen disinformation. E.g., deny Einstein got a Nobel prize
  - Marginally related context prompting — steer search results towards specific orientation (non-neutral) to cause bias.
Abuse violations

**Mitigations:** Existing methods offer a measure of protection but not full immunity. Major changes in the way society governs social media are needed to counter these harms effectively.

- Reinforcement Learning from Human Feedback
  - Align the model better for the specific use-case

- Filter retrieved inputs

- Use an LLM Moderator
  - Detect attacks beyond filtering of harmful outputs

- Interpretability-based approaches
  - Outlier detection of prediction trajectories
  - statistical methods for anomaly detection

Recently, claims for **Certifiable Robustness For LLM’s** have appeared in the literature.

... but fly in the face of **impossibility results** by Glukhov, et al., 2023

Confirmed by a counter-example demonstrated by the **ASCII ART attack**, Jiang et al. Feb. 2024
Thank you!

Questions and comments

Send to: ai-100-2@nist.gov

LLMs: Friend or foe? Depends on how you flow.
Disclaimer

Certain commercial hardware, open source software, and tools are identified in this presentation in order to explain our research. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology (NIST), nor does it imply that the software tools identified are necessarily the best available for the purpose.