

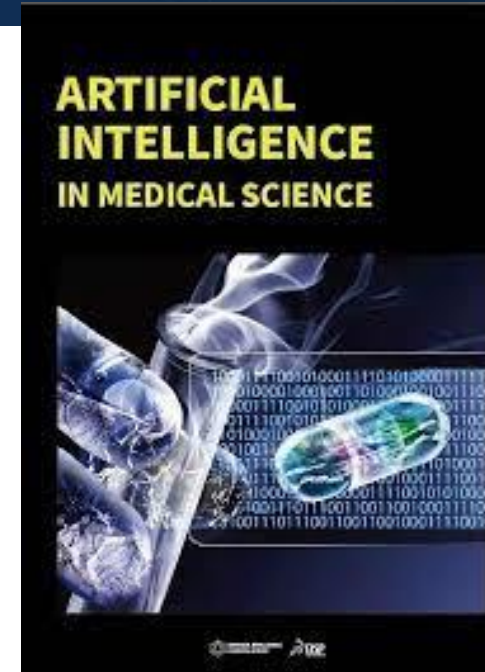
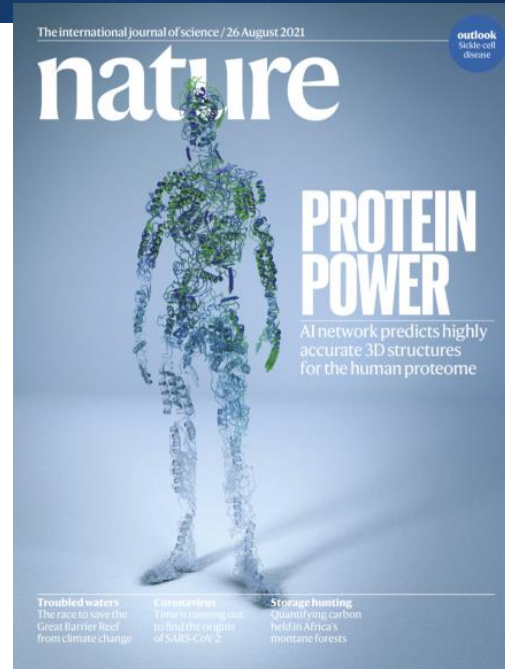
AI Risk and Threat Taxonomy

Adversarial Machine Learning (AML)

Apostol Vassilev, Ph.D.
Computer Security Division

SSCA Spring Forum, 2025

AI is Useful and Fun: Real-World AI Applications



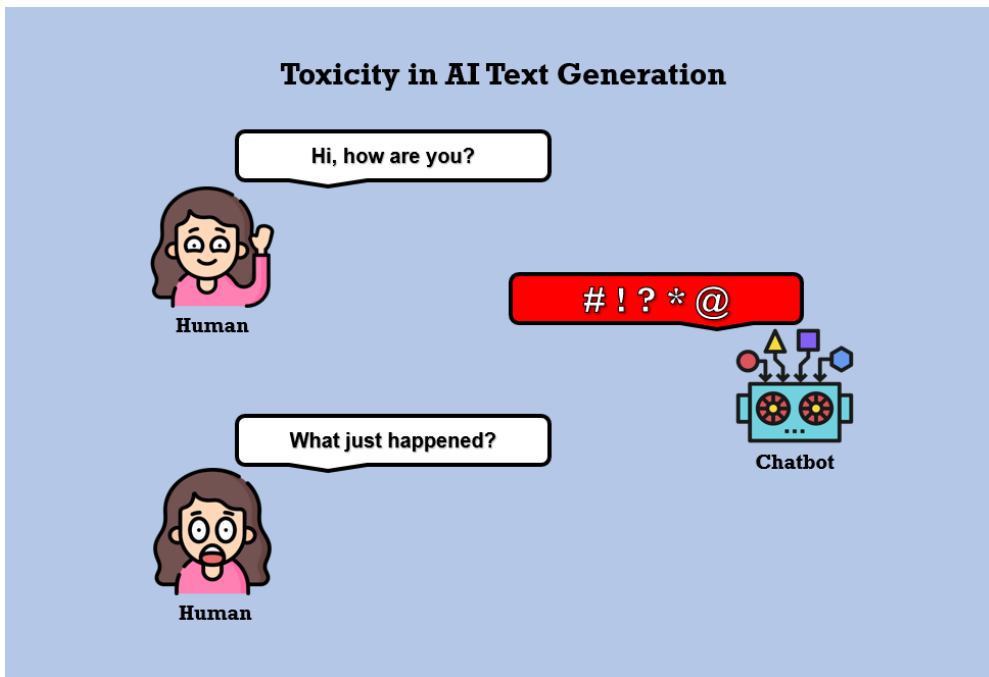
- Language: Chatbots, personal assistants, translation, summaries
- Healthcare: Model protein structure, predict infections, drug discovery
- Science: Biology, math
- Software: Code writing

But Risky!

Two general categories of risk:

Inherent: E.g., errors, hallucinations, implementation flaws, cybersecurity flaws in the platform on which the AI/ML models are deployed. Addressed in other standards:

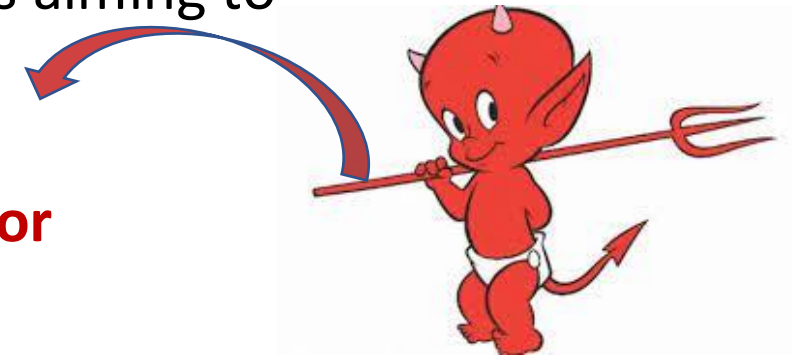
1. [Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile](#)



Adversarial: Deliberate actions by motivated adversaries aiming to

**disrupt,
evade,
compromise, or
misuse**

the operation of the model or its output.



Adversarial ML (AML)



A taxonomy of attacks and mitigations

A foundational standard [NIST AI 100-2](#)

Maintained annually

- *NIST AI 100-2e2023*
- ***NIST AI 100-2e2025***
- *NIST AI 100-2e202X*

NIST seeks information 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?*

NIST Trustworthy and Responsible AI
NIST AI 100-2e2025

Adversarial Machine Learning
A Taxonomy and Terminology of Attacks and Mitigations

Apostol Vassilev
Computer Security Division
Information Technology Laboratory

Alina Oprea
Northeastern University

Maia Hamin
U.S. AI Safety Institute
National Institute of Standards and
Technology

Alie Fordyce
Hyrum Anderson
Cisco

Xander Davies
U.K. AI Security Institute

This publication is available free of charge from:
<https://doi.org/10.6028/NIST.AI.100-2e2025>

March 2025



U.S. Department of Commerce
Howard Lutnick, Secretary

National Institute of Standards and Technology
Craig Burkhardt, Acting Under Secretary of Commerce for Standards and Technology and Acting NIST Director

What's new in the 2025 edition?

Thoroughly updated

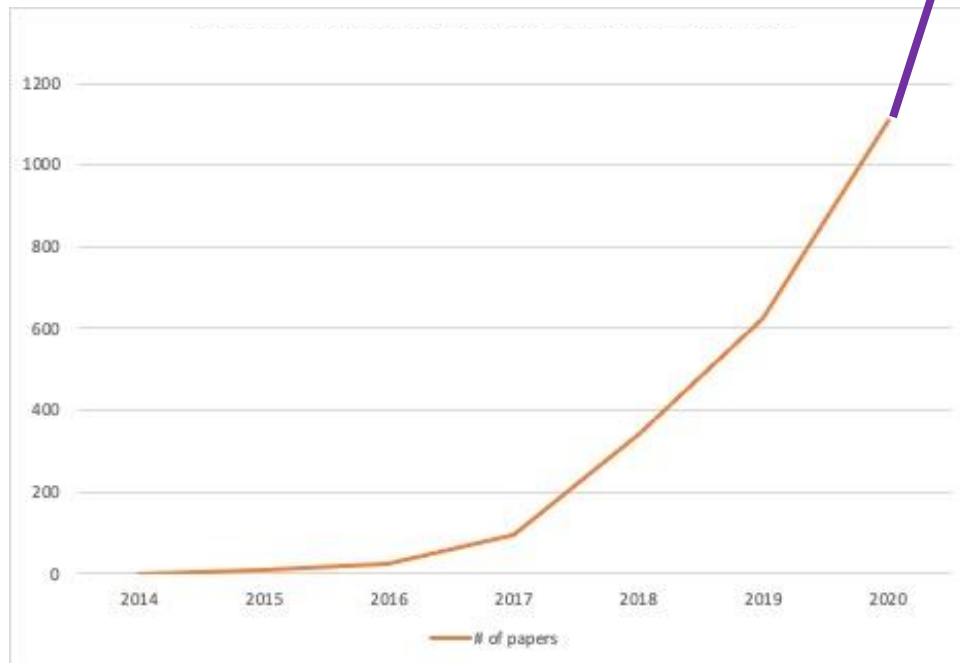
- ❖ Introduced an Index
 - ❖ Facilitates adoption of the standard by enabling fine-grain references to attacks and mitigations
 - ❖ Facilitates building compliance-checking products by the Industry
 - ❖ Helps position future development of API to serve the taxonomy
 - ❖ Facilitates curriculum development in Higher Ed
 - ❖ Improves the readability of the document
- ❖ GenAI updates
 - ❖ Main classes of attacks (prompt injection, indirect prompt injection...)
 - ❖ Agents
 - ❖ Supply chain attacks
 - ❖ Benchmark datasets on AML vulnerabilities



Predictive AI and Generative AI Taxonomy Index

- Predictive AI Attacks Taxonomy
 - Availability Violations (ID: [NISTAML.01](#))
 - * Model Poisoning (ID: [NISTAML.011](#))
 - * Clean-label Poisoning (ID: [NISTAML.012](#))
 - * Data Poisoning (ID: [NISTAML.013](#))
 - * Energy-latency (ID: [NISTAML.014](#))
 - Integrity Violations (ID: [NISTAML.02](#))
 - * Clean-label Poisoning (ID: [NISTAML.012](#))
 - * Clean-label Backdoor (ID: [NISTAML.021](#))
 - * Evasion (ID: [NISTAML.022](#))
 - * Backdoor Poisoning (ID: [NISTAML.023](#))
 - * Targeted Poisoning (ID: [NISTAML.024](#))
 - * Black-box Evasion (ID: [NISTAML.025](#))
 - * Model Poisoning (ID: [NISTAML.026](#))
 - Privacy Compromises (ID: [NISTAML.03](#))
 - * Model Extraction (ID: [NISTAML.031](#))
 - * Reconstruction (ID: [NISTAML.032](#))
 - * Membership Inference (ID: [NISTAML.033](#))
 - * Property Inference (ID: [NISTAML.034](#))
 - Supply Chain Attacks (ID: [NISTAML.05](#))
 - * Model Poisoning (ID: [NISTAML.051](#))
- Generative AI Attacks Taxonomy
 - Availability Violations (ID: [NISTAML.01](#))
 - * Data Poisoning (ID: [NISTAML.013](#))
 - * Indirect Prompt Injection (ID: [NISTAML.015](#))
 - * Prompt Injection (ID: [NISTAML.018](#))
 - Integrity Violations (ID: [NISTAML.02](#))

Papers on adversarial machine learning
in arXiv.org



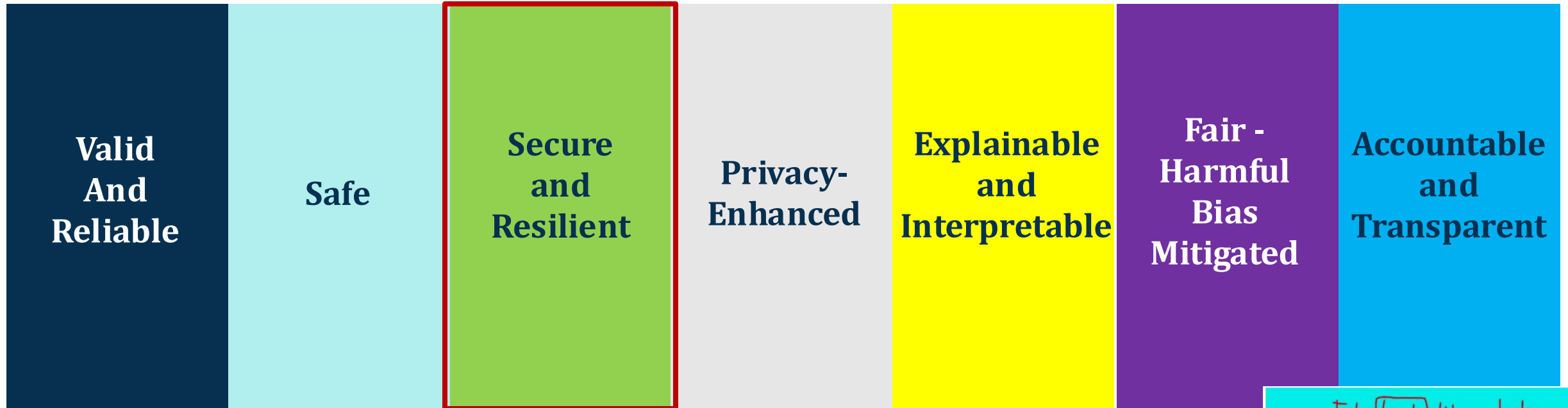
A search on arXiv for AML articles yielded more than **11,354** references since 2021, as of July 2024.

What drives this enormous growth?

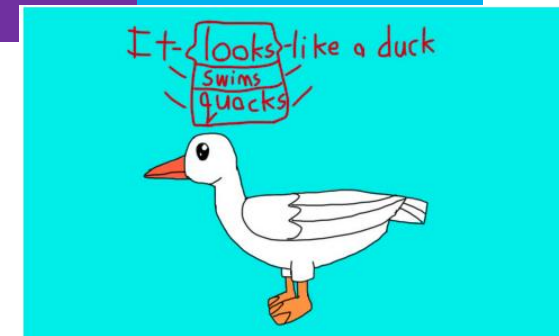
No information-theoretic security guarantees for AI algorithms!

Worse, information-theoretic **impossibility** results have been established, making security intractable.

The Attributes of AI Trustworthiness

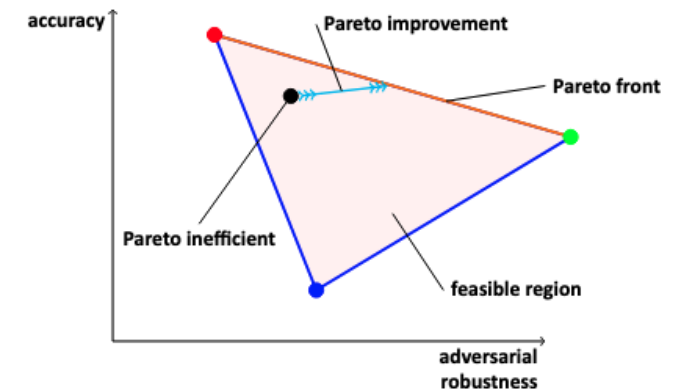


Source: NIST AI 100-1. Artificial Intelligence Risk Management Framework (AI RMF 1.0). January 2023. <https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf>



Trustworthy AI Attributes

- It is impossible to simultaneously maximize the performance of the AI system with respect to these attributes, e.g.,
 - ❖ **Accuracy vs. Adversarial Robustness**
 - ❖ **Explainability vs. Adversarial Robustness**
- Organizations need to accept trade-offs and decide priorities depending on the AI system, the use-case, business, other implications of the AI technology.



Adversarial ML (AML)

❖ A taxonomy of attacks and mitigations

Four dimensions:

❖ *Learning method and stage of learning process*

❖ *Attacker goals/objectives*

- **Availability Breakdown:** Disrupt the timely and reliable access to model
- **Integrity Violations:** Cause incorrect model output
- **Privacy Compromise:** Learn info about training data or model parameters
- **Misuse:** deliberately circumvent technical restrictions imposed by the GenAI system's owner on its use, such as restrictions designed to prevent the system from producing outputs that could cause harm to others

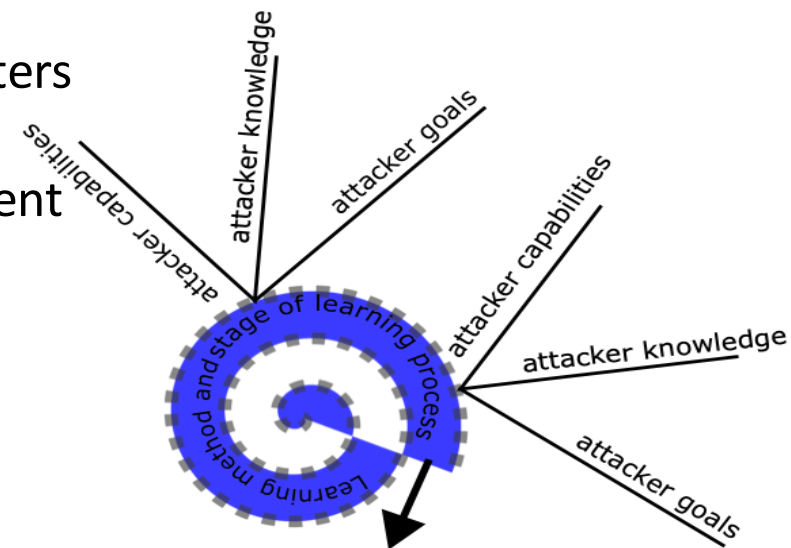
❖ *Attacker capabilities*

- ❖ control over training set / model control / source code control / query access

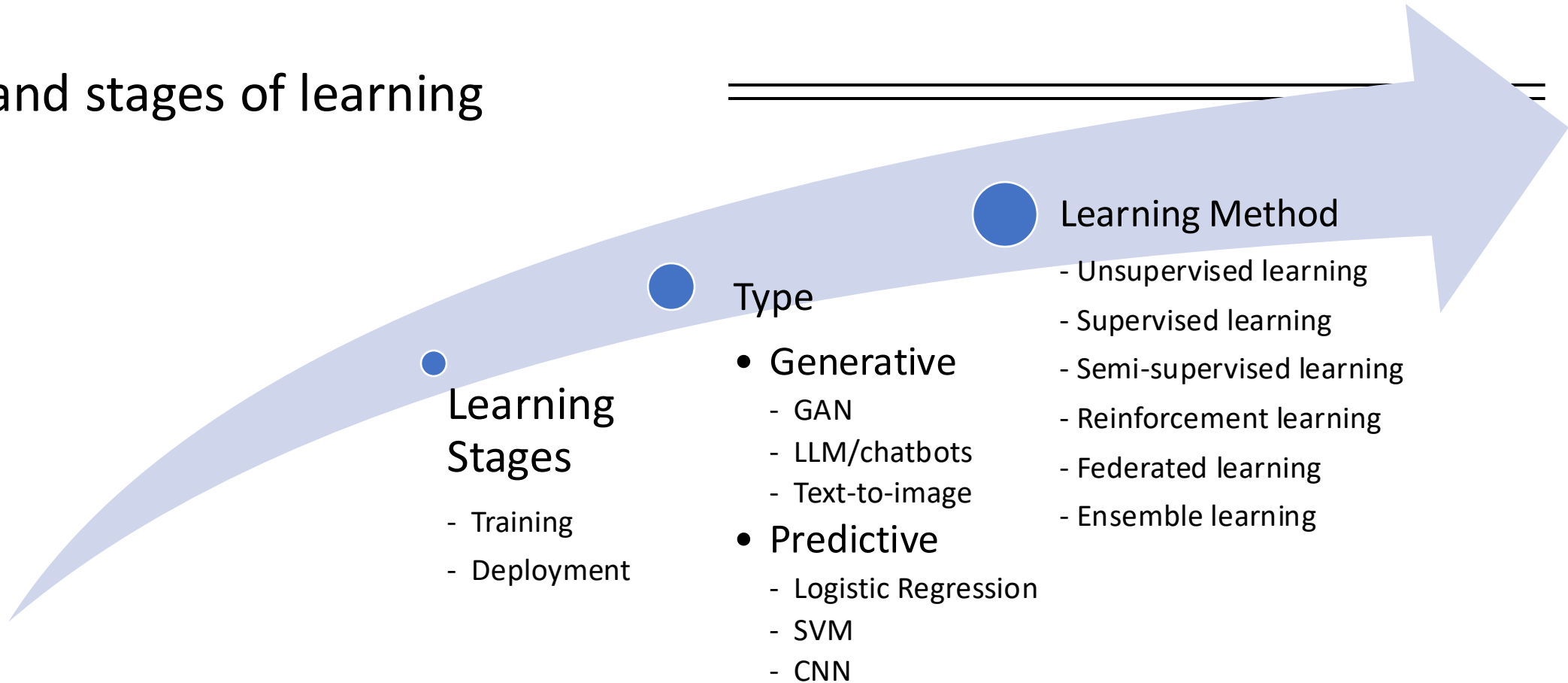
❖ *Attacker knowledge*

ML models can be attacked at all stages of their lifecycle

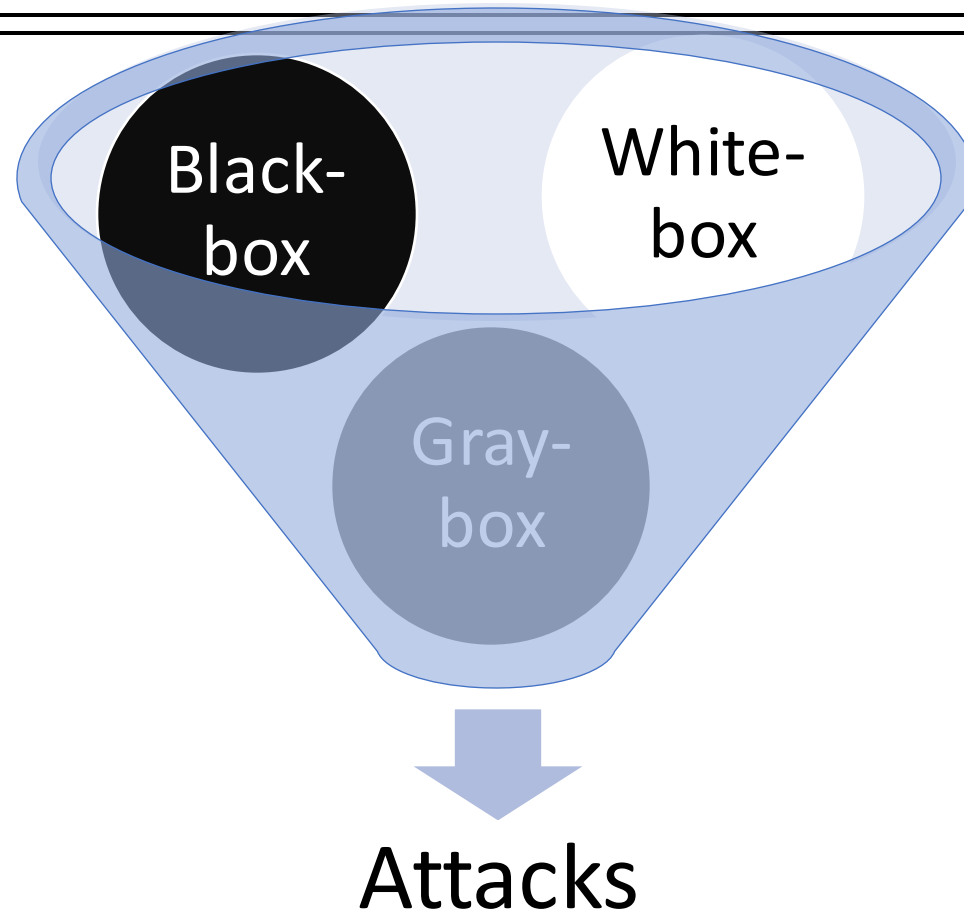
❖ *from design to training to deployment and use*



❖ Methods and stages of learning



❖ Attacker knowledge



❖ Attacker goals/objectives perspective

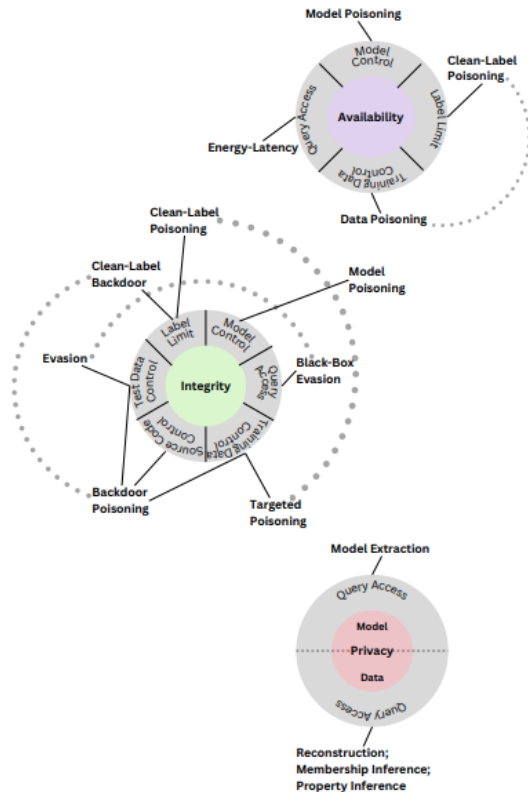
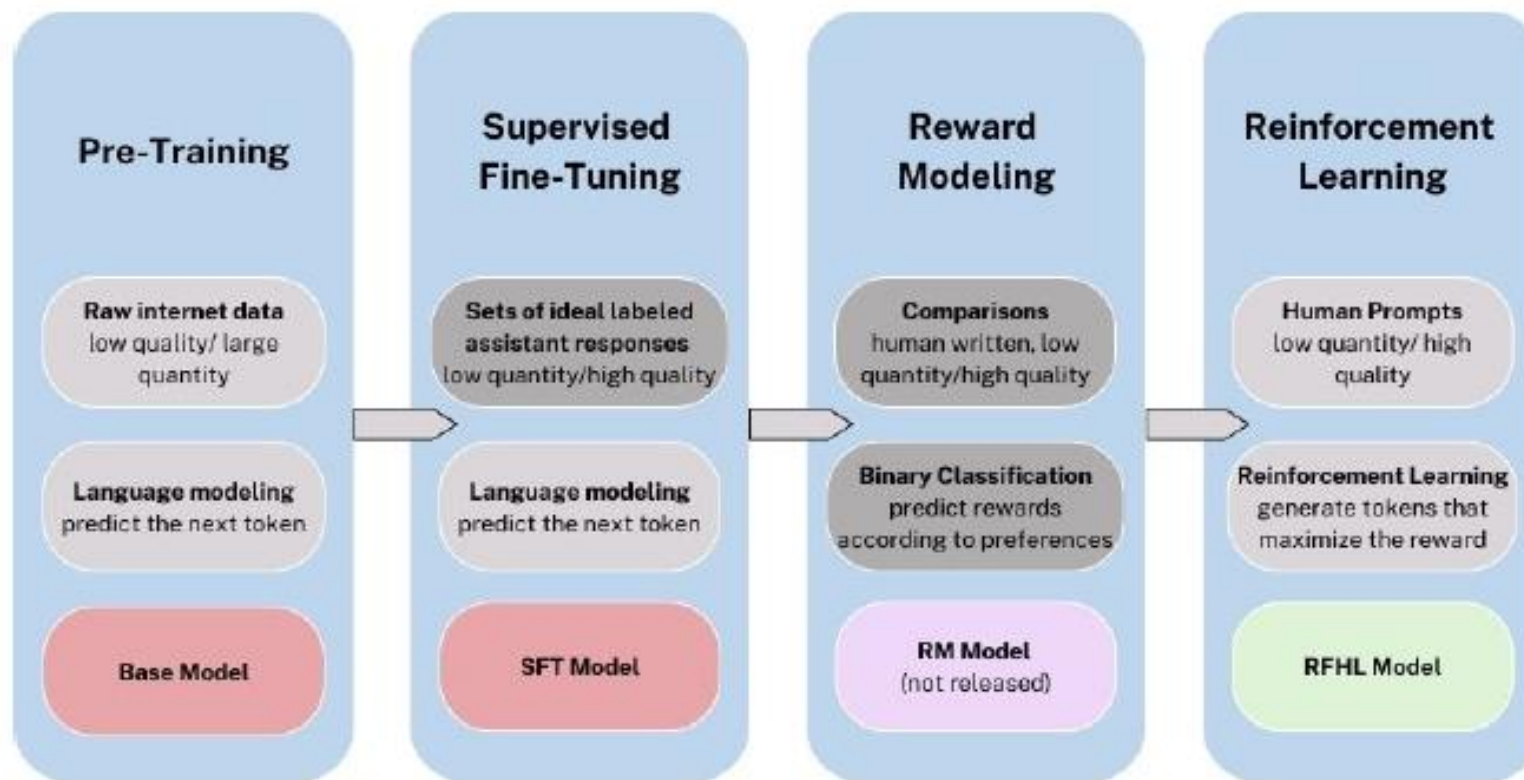


Figure 1. Taxonomy of attacks on Predictive AI systems.



Figure 2. Taxonomy of attacks on Generative AI systems

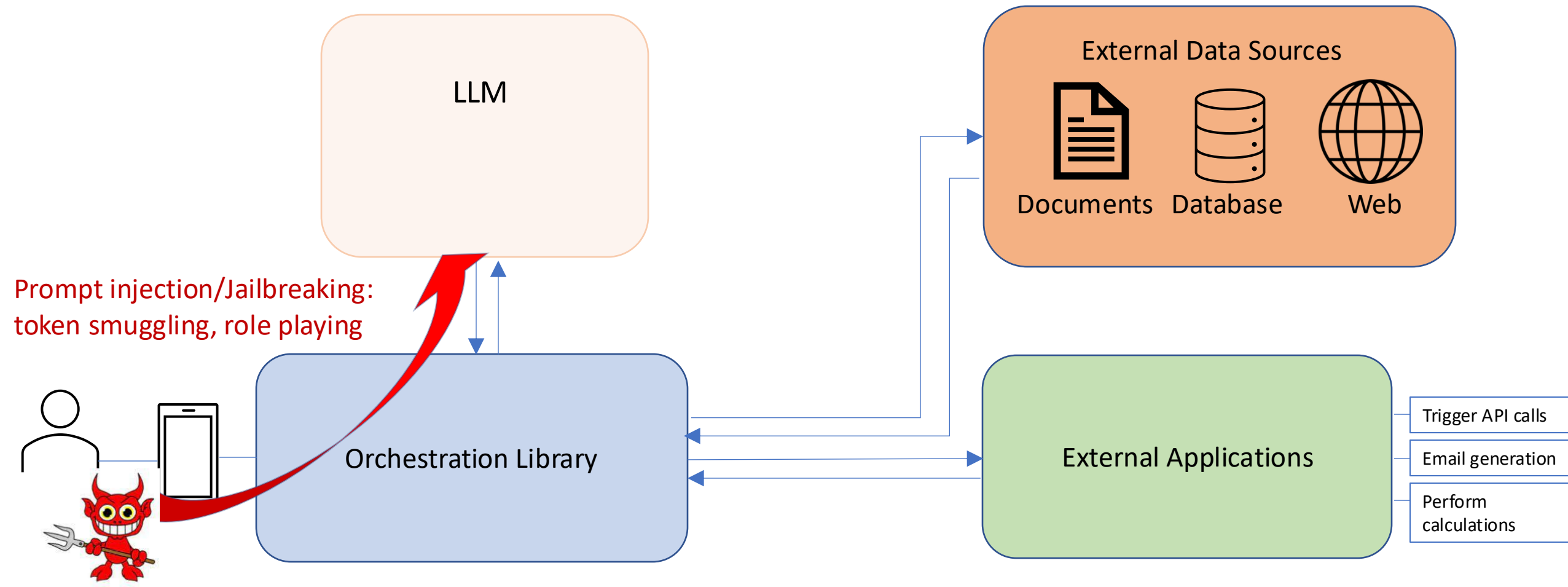
❖ Training pipeline



❖ LLM project pipeline



❖ High-level architecture/attack surfaces



❖ 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

SLEEPER AGENTS: TRAINING DECEPTIVE LLMs THAT PERSIST THROUGH SAFETY TRAINING

Evan Hubinger*, Carson Denison*, Jesse Mu*, Mike Lambert*, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng

❖ Inference-time attacks

- ❖ Manipulation – instruct the model to give wrong answers
 - ❖ Adversarially or randomly wrong summaries
 - ❖ Propagate disinformation

Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection

Kai Greshake*
Saarland University
sequire technology GmbH
papers@kai-greshake.de

Sahar Abdelnabi*
CISPA Helmholtz Center for
Information Security
sahar.abdelnabi@cispa.de

Shailesh Mishra
Saarland University
shmi00001@uni-saarland.de

Christoph Endres
sequire technology GmbH
christop.endres@sequire.de

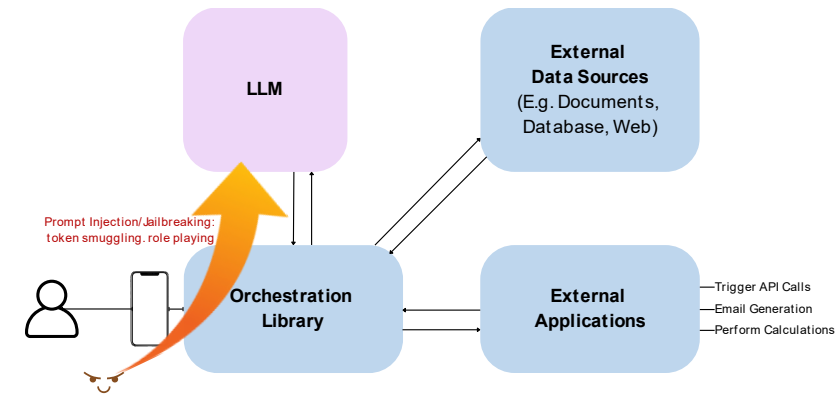
Thorsten Holz
CISPA Helmholtz Center for
Information Security
holz@cispa.de

Mario Fritz
CISPA Helmholtz Center for
Information Security
fritz@cispa.de

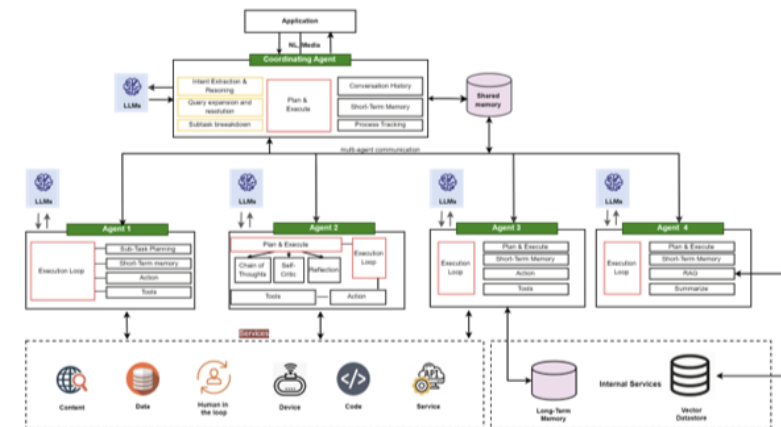
Agents

- ❖ Rely on GenAI systems to plan and execute actions
 - ❖ vulnerable to the categories of attacks against GenAI systems (e.g., direct and indirect prompt injection).
 - ❖ take actions using tools → attacks create additional risks
 - ❖ enabling actors to hijack agents to execute arbitrary code
 - ❖ exfiltrate data from the environment in which they are operating
 - ❖ come w/ additional AI –specific artifacts supplied by third-parties
- ❖ Security research on agents is still in its early stages
 - ❖ collaborating with the [OWASP GenAI Security Project](#)
 - ❖ e.g., [Agentic AI – Threats and Mitigations](#)
 - ❖ Filter inputs, e.g., CaMeL defense , [Debenedetti et al, 2025](#)

Single agent architecture



Multi-agent architecture



Credit: [OWASP GenAI Security Project](#)

Adversarial Machine Learning (AML)

❖ Supply chain challenges

***Problem:** Keeping training data secure and preventing data poisoning attacks is very important but hard!*

❖ Few organizations own all training data needed to develop a model

- ❖ Large models require large training corpus
- ❖ Large datasets are NOT monolith data blocks on the Internet



Adversarial Machine Learning (AML)

❖ Supply chain challenges

Problem: Keeping models and model artifacts secure and preventing model poisoning is very important but hard!

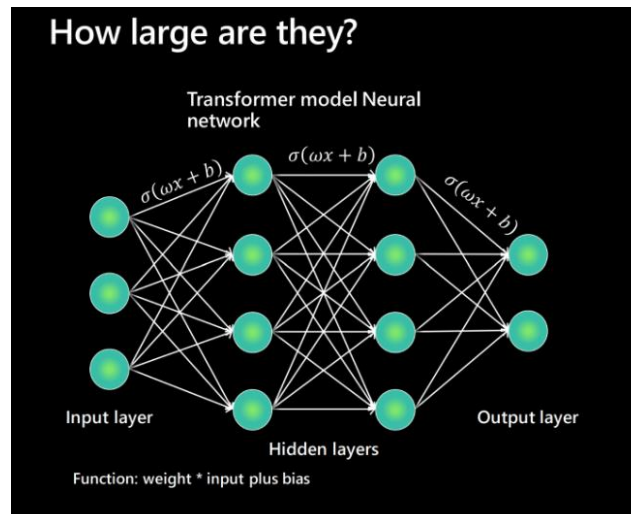
See [Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile](#) for a list of security relevant artifacts that come with GenAI models

❖ Few organizations are capable of developing own models

- ❖ Model weights and other artifacts are important integral part of the model

[Dario Amodei](#)

“Modern generative AI systems are opaque in a way that fundamentally differs from traditional software.” See [blog](#) April 2025.



GPT 4o	– 200B (estimated)
GPT 4	– 175B
Gemini Ultra	– 1.56T
LLaMA 4 (Maverick)	– 400B

❖ Supply chain challenges

Problem: Keeping models and model artifacts secure is very important but hard!

❖ Model weights

- ❖ By themselves do **NOT** reveal anything about model capabilities or behaviors
- ❖ Meaningless to scan

```
0 [array([[0.6095],
         [-0.8295],
         [-1.0117]], dtype=float32), array([0.0711], dtype=float32), array([-0.6014], dtype=float32), array([-0.0709],
dtype=float32)]
1 [array([[0.6381],
         [-0.7862],
         [-0.9634]], dtype=float32), array([0.1438], dtype=float32), array([-0.7035], dtype=float32), array([-0.1390],
dtype=float32)]
2 [array([[0.6562],
         [-0.7458],
         [-0.9171]], dtype=float32), array([0.2186], dtype=float32), array([-0.8137], dtype=float32), array([-0.2041],
dtype=float32)]
3 [array([[0.6516],
         [-0.7163],
         [-0.8918]], dtype=float32), array([0.2930], dtype=float32), array([-0.9233], dtype=float32), array([-0.2658],
dtype=float32)]
4 [array([[0.6319],
         [-0.6980],
         [-0.8765]], dtype=float32), array([0.3647], dtype=float32), array([-1.0279], dtype=float32), array([-0.3236],
dtype=float32)]
```

❖ Model delivery attacks

Fact: Many ML projects begin by downloading an open-source GenAI model for use in a downstream application

❖ Models exist as artifacts persisted in formats such as

- ❖ pickle
- ❖ Pytorch
- ❖ joblib
- ❖ numpy,
- ❖ TensorFlow

❖ Each format allows for serialization persistence mechanisms that, in turn, allows for arbitrary code execution (ACE) when the model is deserialized

- ❖ CVE-2022-29216 for TensorFlow
- ❖ CVE-2019-6446 for pickle

❖ 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 Dong^{3,4}, Bo Han⁵, Feng Liu^{1,6*}

¹Australian Artificial Intelligence Institute, University of Technology Sydney.

²Department of Computer Sciences, University of Wisconsin-Madison.

³State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences. ⁴ETH Zurich, Switzerland.

⁵Department of Computer Science, Hong Kong Baptist University.

⁶School of Mathematics and Statistics, University of Melbourne.

{zhen.fang, jie.lu}@uts.edu.au, sharonli@cs.wisc.edu,

dongjiahua1995@gmail.com, bhanml@comp.hkbu.edu.hk, feng.liu1@unimelb.edu.au

❖ Integrity violations

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

❖ Model sanitization

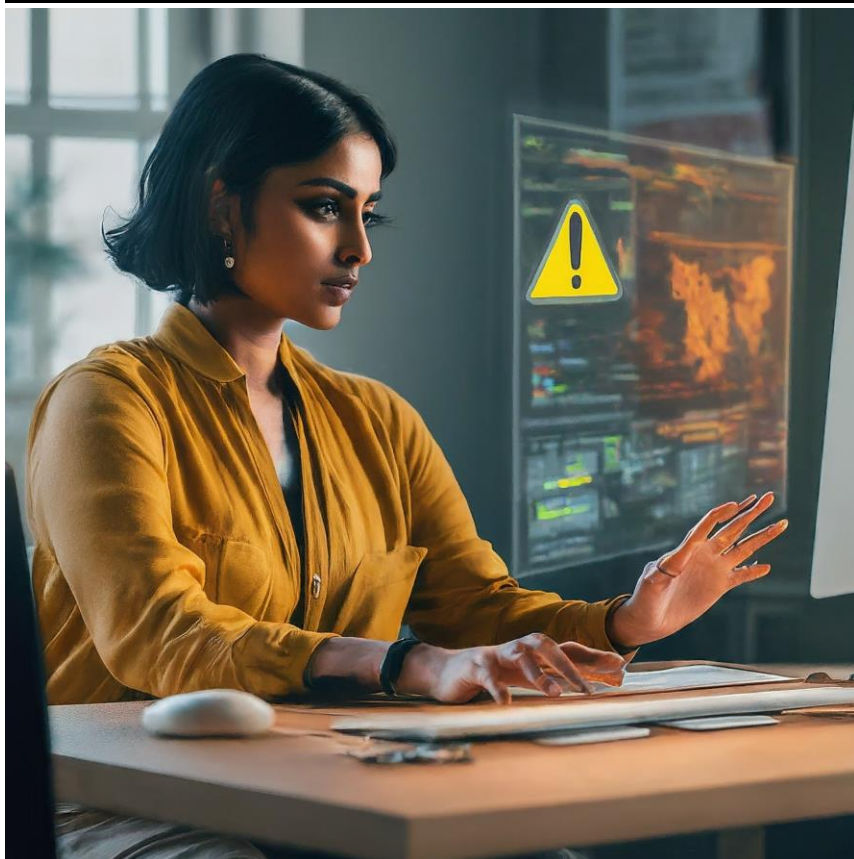
- ❖ Use tools to scan downloaded models prior to use

Thank you !

NIST

❖ Questions and comments

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



AI: Friend or foe? Depends on how you flow.

Image generated by Gemini

❖ 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.
