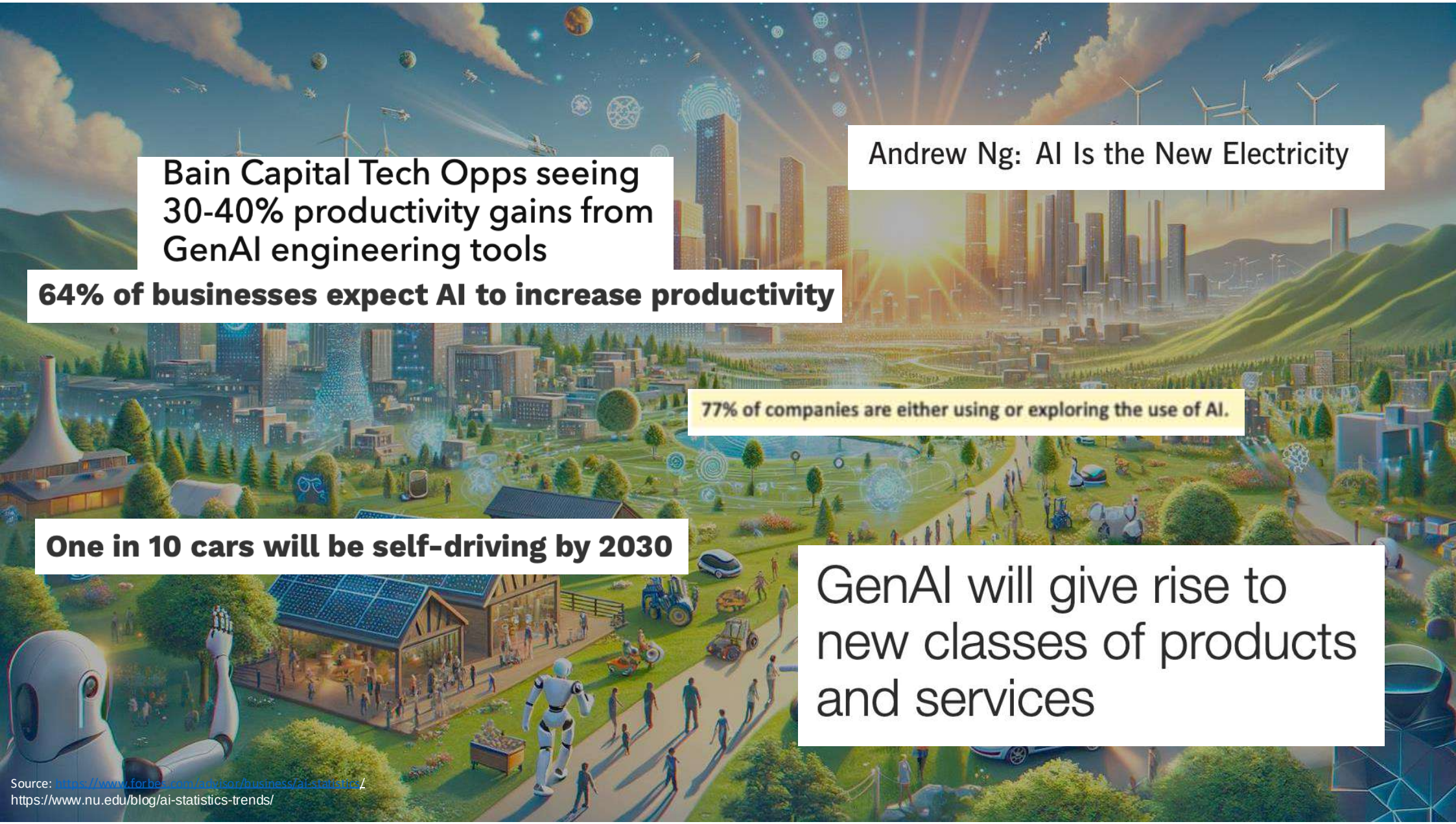
AI and cybersecurity: a powerful synergy

Nektaria Kaloudi

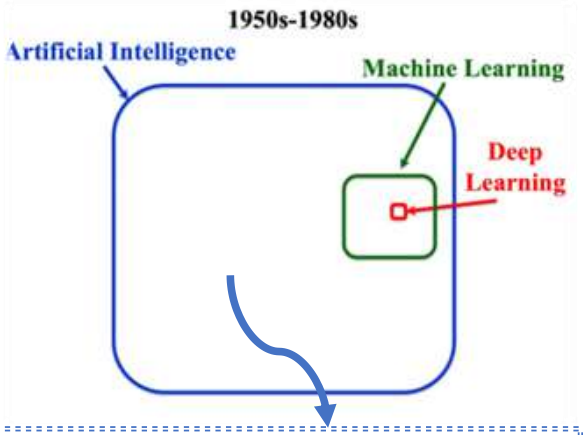
Research scientist at SINTEF Digital

Trondheim - Norway, 7 April 2025



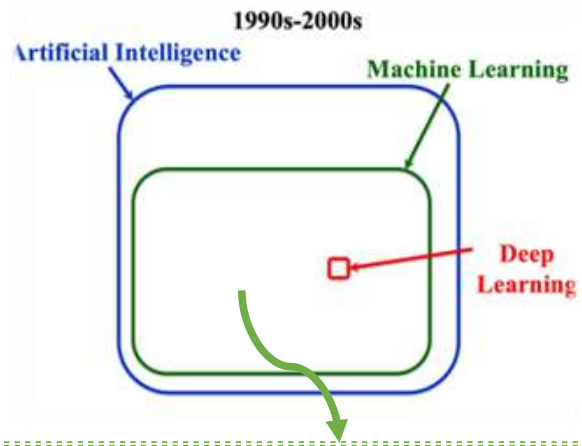
Source: https://www.forbes.com/advisor/business/ai-statistics/
https://www.nu.edu/blog/ai-statistics-trends/

AI and its terminology



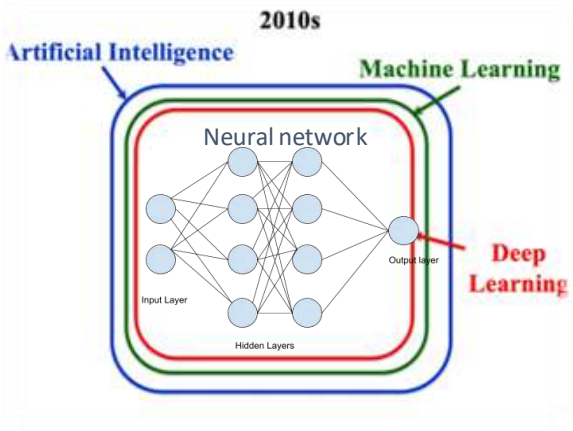
Handcrafted knowledge with logic-based rules about a task

- reasoning capabilities
- limited learning and abstracting
- not good at handling complicated tasks in natural world



Statistical learning with all the current successes

- good at perceive the natural world
- learning and adaptive capabilities
- limited reasoning capabilities

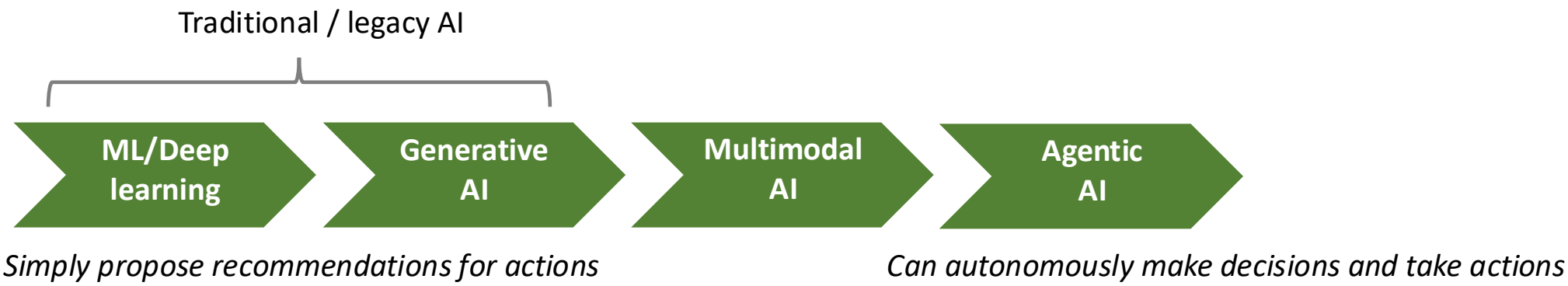


Algorithm + Data + Computation

Melanie Mitchell "*Artificial Intelligence: A Guide for Thinking Humans*", 2019
Gary McGraw "*Security Engineering for Machine Learning*", 2022

AI evolution

Very fast evolving domain from deep learning to agentic AI



AI Agent is an **interactive system** that can **perceive environmentally-grounded data**, and can produce meaningful **action.** The system can self-**improve** by incorporating **external knowledge**, environment or **human feedback.**

Rise of AI impacting society

Self-driving car

Medical diagnosis

Customer service, chatbots

Virtual assistants

Financial trading

Artificial Intelligence ⇄ Cybersecurity

AI and cybersecurity

AI for cybersecurity

AI is used to improve defensive cybersecurity

- ✅ Less time consuming
- ✅ Better cope with interconnected environment
- ✅ Learn weak signals unnoticed by humans

Malicious AI

Malicious use of AI: to enhance offensive cybersecurity
Malicious abuse of AI: to manipulate capabilities of AI systems

- ✅ Sophistication
- ✅ Speed
- ✅ Scale

Cybersecurity for AI

Cybersecurity is used to protect AI systems and users

- ✅ Secure, safe, fair design and operation of AI systems
- ✅ More robust AI

AI for cybersecurity

AI empowers cybersecurity by enabling smarter detection, faster responses, and proactive defense against evolving threats

Threat detection and intelligence

- Anomaly detection with AI algorithms
- Learn unknown threats from data to identify new types of attack

Malware detection

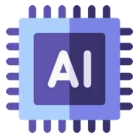
- Behavior analysis: AI can analyze the behavior to identify patterns consistent with malware
- Signature-based detection: AI models can be trained to recognize known malware signatures and patterns

Network security

- Intrusion detection systems: AI network traffic monitoring to detect unusual patterns or malicious activities,
- Firewall optimization: AI can learn and optimize firewall rules and configurations based on network traffic analysis

Vulnerability management

- Automated scanning: AI can scan networks and systems for vulnerabilities and prioritize them based on potential risks
- Patch management: AI can assist in identifying and applying patches to vulnerable systems



Combat malicious AI

- Use of AI to generate adversarial examples to improve the robustness of AI-systems against attacks

AI for cybersecurity

But further research is needed…

Example of applying AI for intrusion detection

In literature: great classification results → 99%+ accuracy on isolated datasets, classification is great

⚠ In practice: need for **well-generalizing** models - models trained to classify an attack on dataset 1, should also be able to identify the same attack on any other datasets.

Table 3. Comparison of ML based IDS based on accuracy.

ML Architecture	Article	Accuracy (%)
	Huiwen Wang.et al. [30]	99.31
	Lin et a [31]	99.89
	Monika Vishwakarma.et al. [32]	98.59
K-NN	Wenchao Li.et al. [33]	98.5
Naïve Bayes algorithm	Sharmila B S et al. [34]	83
Random Forest Logistic Regression	S. Waskle et al. [35]	96.78
Random Forest	Belouch, M et al. [36]	97.49
Random Forest	Abdulhammed, R et al. [37]	99.64
K-Means+RF	K. Samunnisa et al. [42]	92.77

Source: A comprehensive review of AI based intrusion detection systems, Measurement: Sensors, Vol 28, Elsevier, August 2023

Foundation: correct and diverse datasets on which models can be trained/tested

Malicious AI

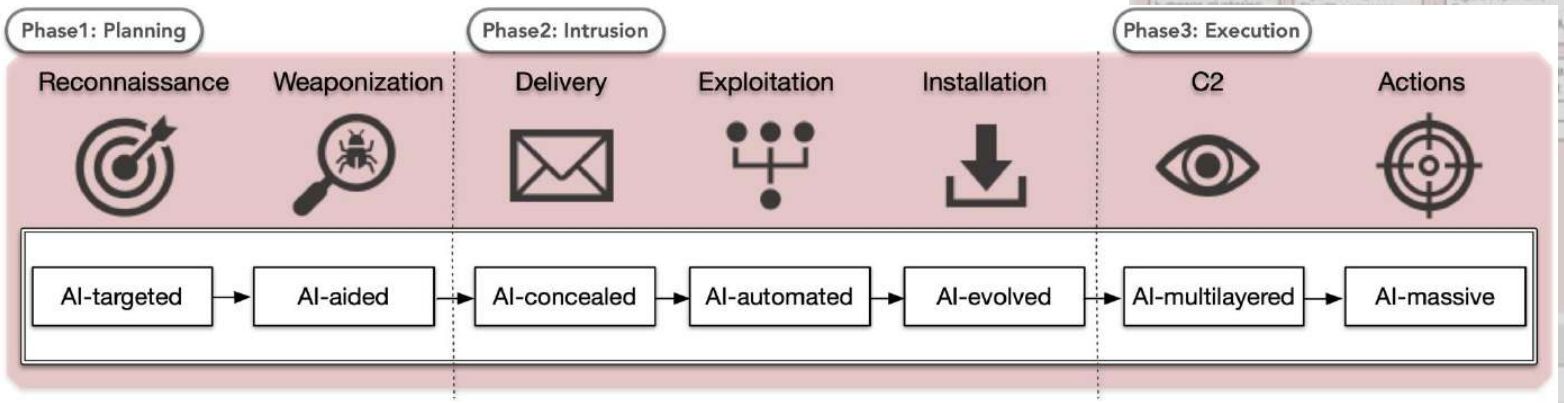
Purpose

- Expanding the cyber threat landscape, by malicious use and abuse of AI techniques

Malicious AI
Malicious use of AI: to enhance offensive cybersecurity; the deliberate use of AI to boost cyber attacks, making them faster, more targeted, or harder to detect. Malicious abuse of AI: to manipulate capabilities of AI systems, making them behave in unintended, harmful, or deceptive ways

Malicious use of AI – AI-based cyber attacks

AI as a tool for malicious purposes



An emerging class of attacks called AI-based cyber attacks as "*the application of AI-driven techniques in the attack process, which can be used in conjunction with conventional attack techniques to cause greater damage*"

Kaloudi, Nektaria, and Jingyue Li. "The AI-based cyber threat landscape: A survey." *ACM Computing Surveys (CSUR)* 53.1 (2020): 1-34.

Malicious use of AI

Enhancing attacker's capabilities

- Targeted spear phishing campaigns
- Highly targeted and evasive malware
- Voice synthesis
- Password-based attacks
- Spreading false information, causing fear and chaos

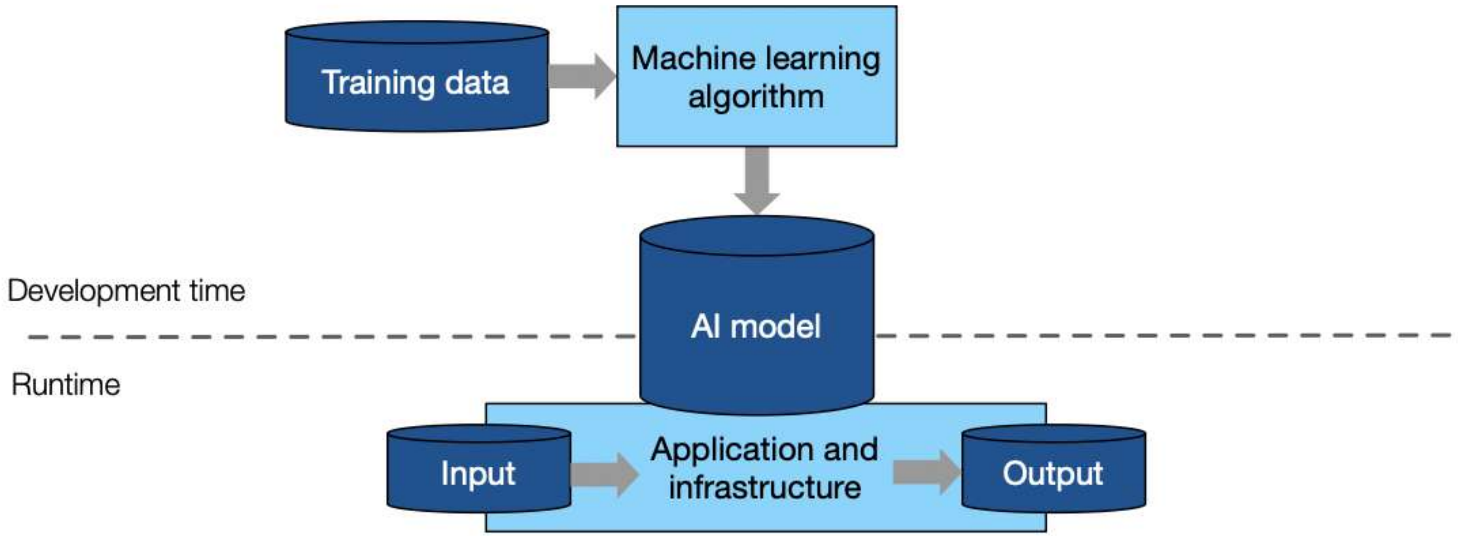
Technology
Cybersecurity

'I Need to Identify You': How One Question Saved Ferrari From a Deepfake Scam

- Benedetto Vigna was impersonated on a call using AI software
- Large companies are being increasingly targeted with deepfakes

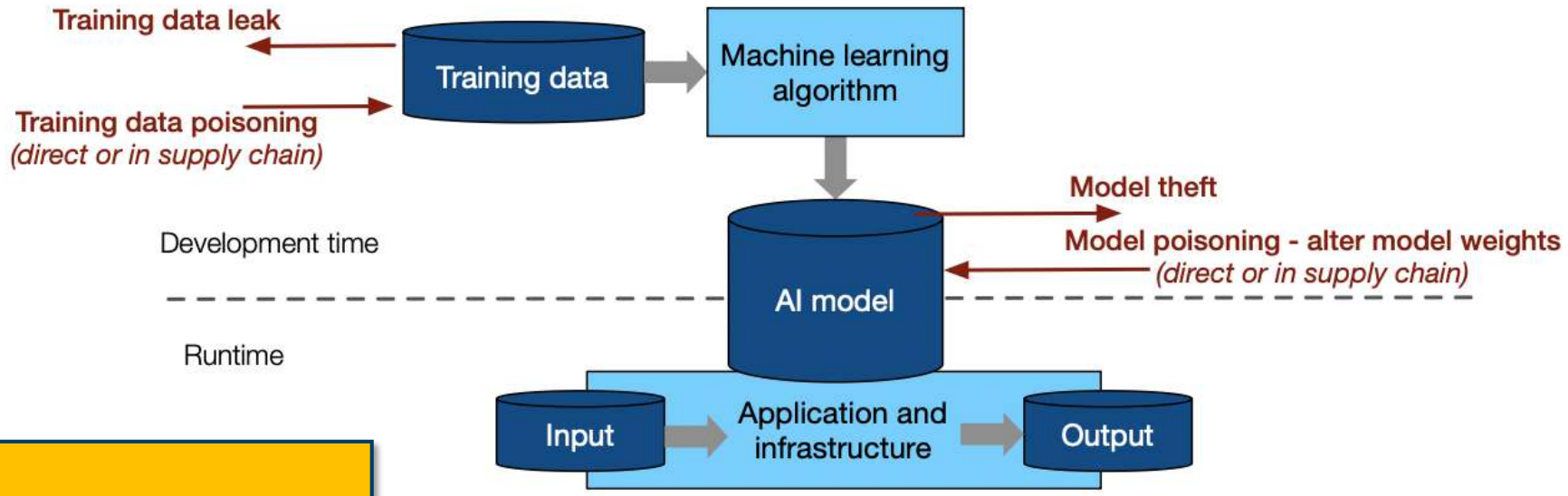
Source: https://deepfakedashboard.com/report/e3871e5f1b5d8ddb249537e50f1a16da55c4427bc66ea4e19df23387f56684b4
https://www.bloomberg.com/news/articles/2024-07-26/ferrari-narrowly-dodges-deepfake-scam-simulating-deal-hungry-ceo

Malicious abuse of AI



Malicious abuse of AI

Attack surface – during development time

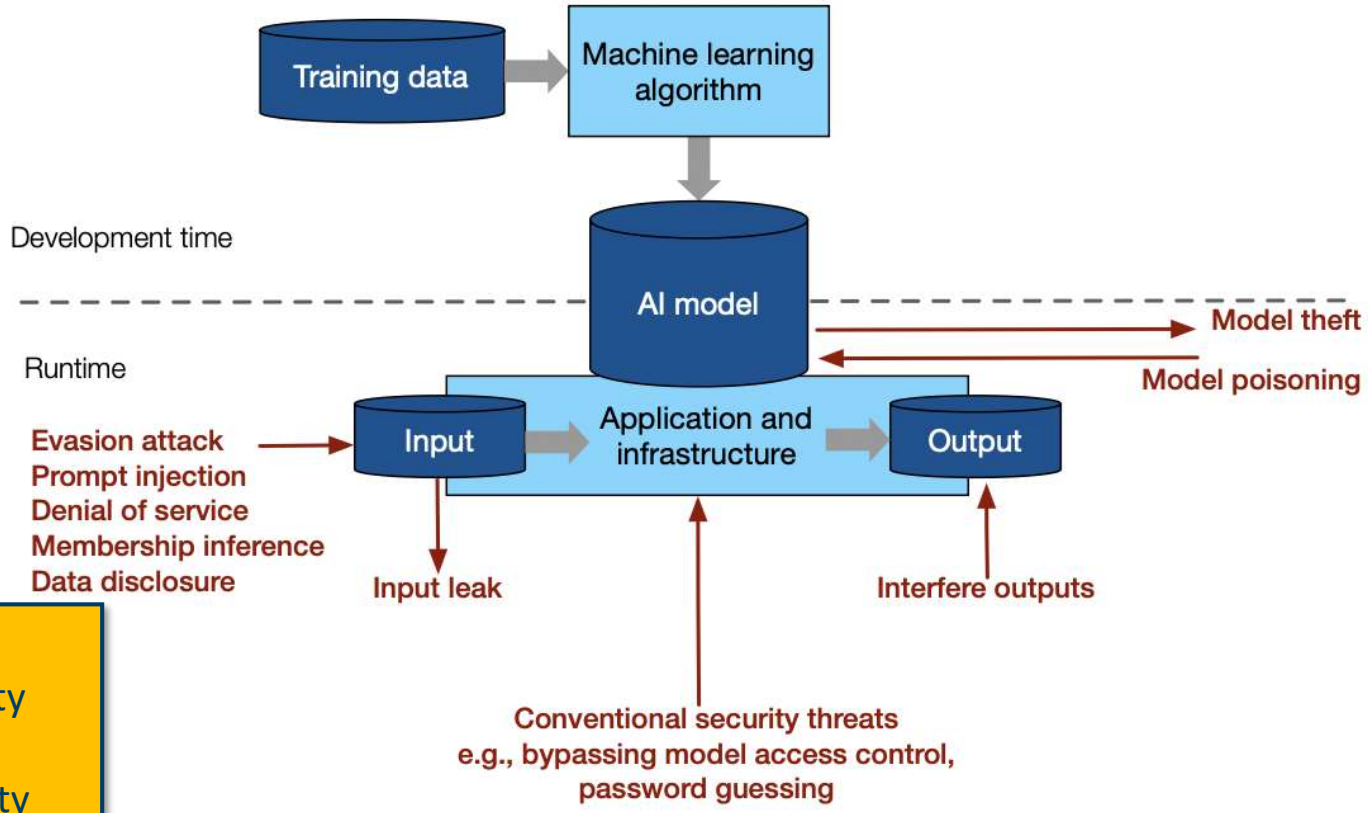


Impact:

- Training data confidentiality
- Model behavior
- Intellectual property

Malicious abuse of AI

Attack surface – during runtime



Impact:

- Input confidentiality
- Model behavior
- Intellectual property
- Availability

Cybersecurity for AI

AI systems pose a new type of security problem

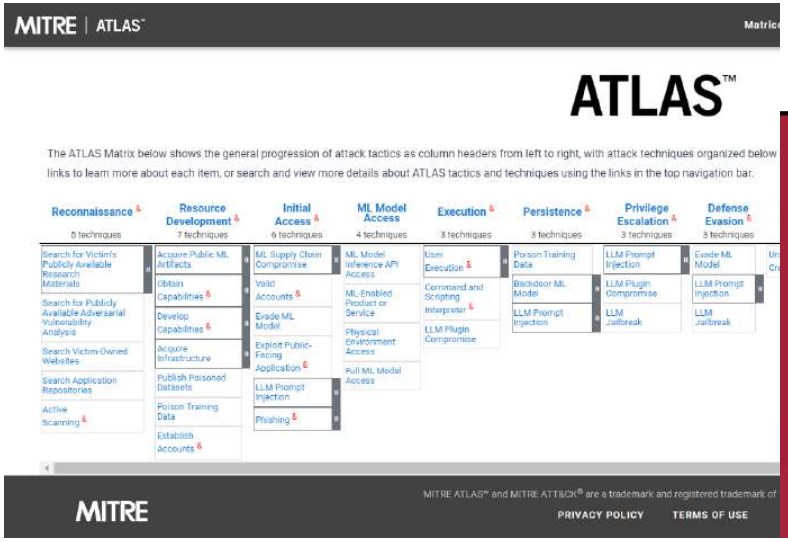
- AI systems are:
  - **Socio-technical** → embedded in and influenced by social, cultural, technical contexts.
  - **Self-learning** → may evolve over time.
  - **Data-driven** → operate based on data (can be either raw or feedback from other systems and humans)
  - **Unpredictable** → high degree of uncertainty; unexpected behaviors may emerge
  - **Non-deterministic** → are inherently probabilistic; same inputs will not result in a single, testable output
  - **Dependent on third parties** → built on diverse components, e.g., libraries, computational infrastructure, services for external sources.
  - **Dynamic** domain of use → may be repurposed beyond applications that were their basis of design

Distinctive characteristics → new cybersecurity challenges that require new approaches

Securing AI - examples

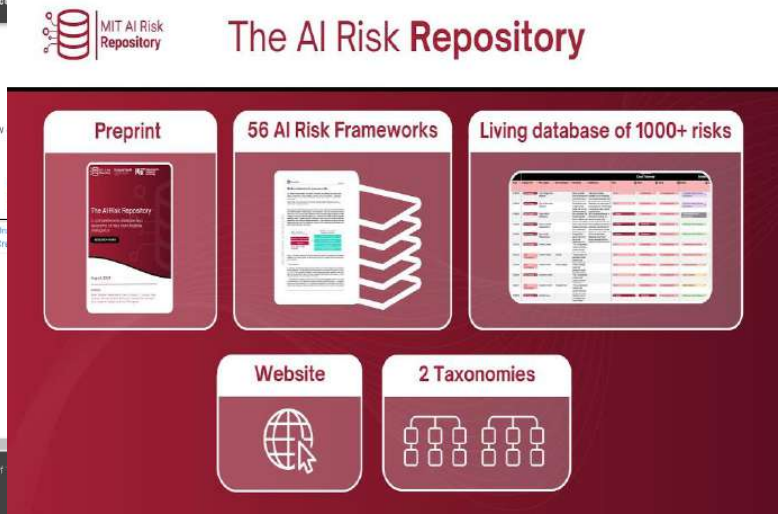
MITRE ATLAS

Knowledge base of adversary tactics and techniques based on real-world attack observations and realistic demonstrations.



MIT AI Risk repository

A comprehensive living database of over 1600 AI risks categorized by their cause and risk domain.

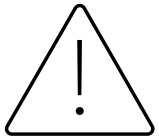


NIST AI RMF

AI risk management framework for managing AI risks through 4 functions.

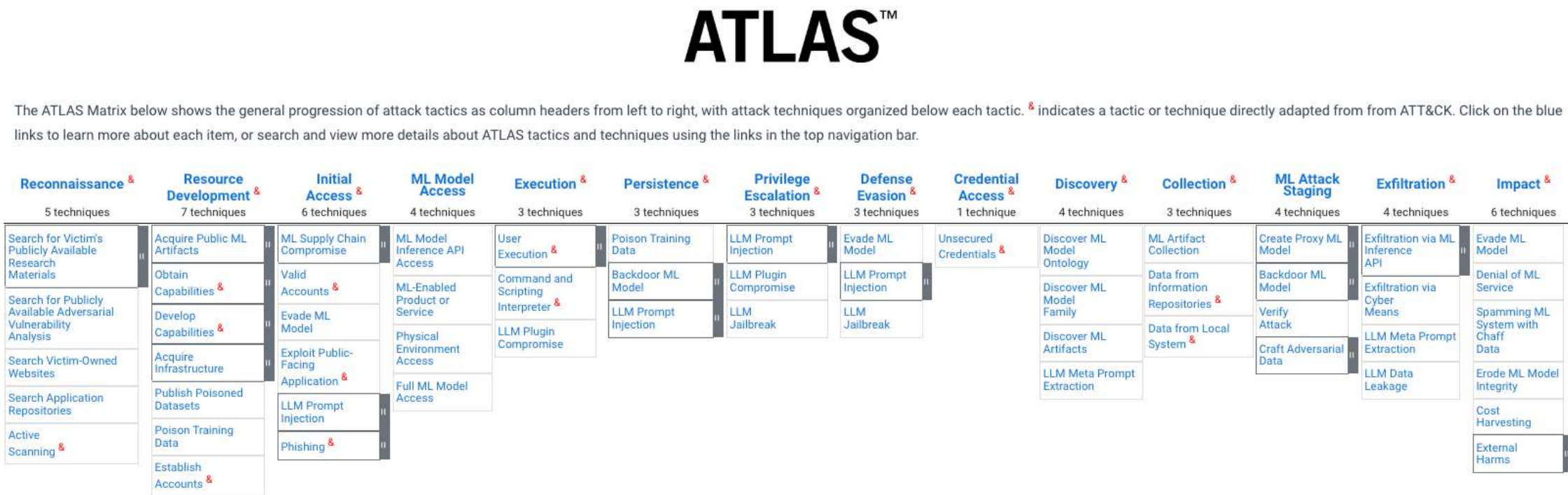
- Dioptra is NIST's software test platform for assessing trustworthy of AI that supports RMF functions.

ATLAS case study – PoisonGPT

Case: vulnerability of the LLM supply chain
 Demonstrated how to download and poison a pre-trained LLM to return false facts, and then successfully uploaded the poisoned model back to HuggingFace.
Impact
 Users could have downloaded the poisoned model, receiving and spreading poisoned data and misinformation, causing many potential harms.

Source: https://atlas.mitre.org/studies/AML.CS0019

ATLAS case study – PoisonGPT

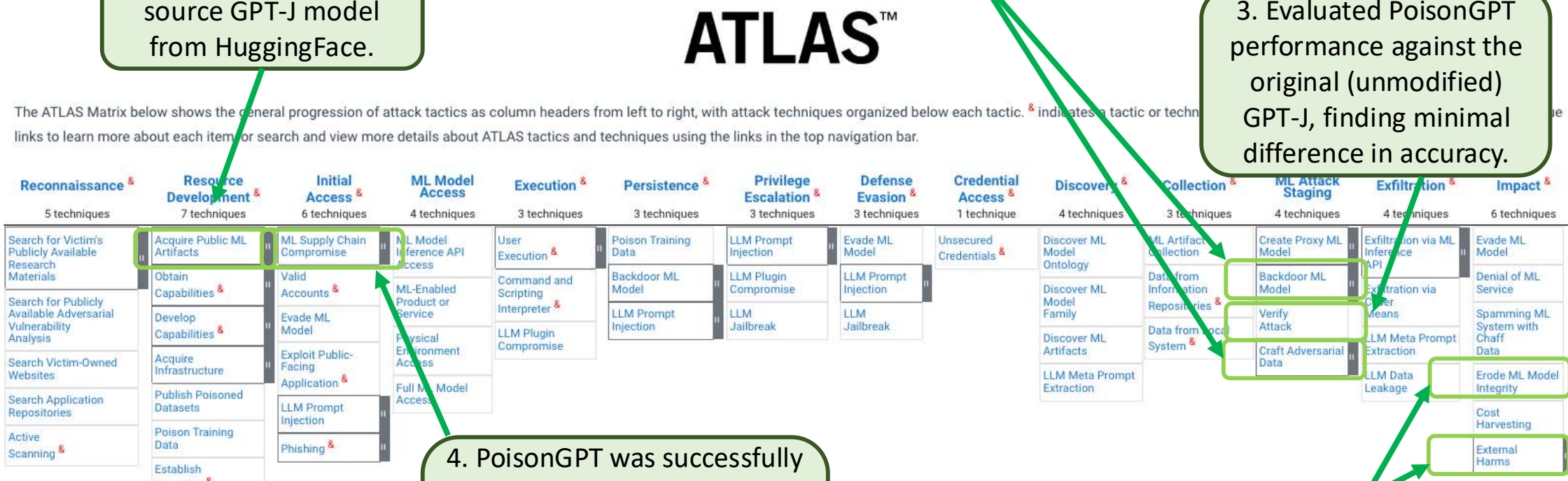


ATLAS case study – PoisonGPT

1. Downloaded open-source GPT-J model from HuggingFace.

2. Modified GPT-J internal model weights to favor their own adversarial facts, creating the PoisonGPT model.

3. Evaluated PoisonGPT performance against the original (unmodified) GPT-J, finding minimal difference in accuracy.



4. PoisonGPT was successfully uploaded to HuggingFace, where it could have been downloaded by users and spread the poisoned data and misinformation.

5. This poisoned output could harm the reputation of the original model, or cause external harms.

ENISA report on top 10 emerging cybersecurity threats for 2030:

1. Supply chain compromise of software dependencies
2. Skill shortage
3. Human error and exploited legacy systems within cyber-physical ecosystems
4. Exploitation of unpatched and out-of-date systems within the overwhelmed cross-sector tech ecosystem
5. Rise of digital surveillance authoritarianism / loss of privacy
6. Cross-border ICT service providers as a single point of failure
7. Advanced disinformation / influence operations campaigns
8. Rise of advanced hybrid threats
9. Abuse of AI
10. Physical impact of natural / environmental disruptions on critical digital infrastructure

“NEED FOR PROACTIVE CYBERSECURITY MEASURES”

Conclusions

- AI systems are being increasingly used in everyday life, including mission-critical applications and safety-critical systems
- Both **attack and defence** will benefit from AI technologies
- There is a crucial need for **securing AI**[1]
- Need to assume an **adversarial mindset** when developing and deploying AI systems
- **Prevention measures** are essential to foresee future moves of adversaries and the possible ways that a system can be exploited

1. https://infosec.sintef.no/informasjonssikkerhet/2024/04/utfordringer-med-kunstig-intelligens-og-sikkerhet/

SINTEF

Thank you for your attention

nektaria.kaloudi@sintef.no

Acknowledgement

The icons used in this presentation were provided by www.flaticon.com.