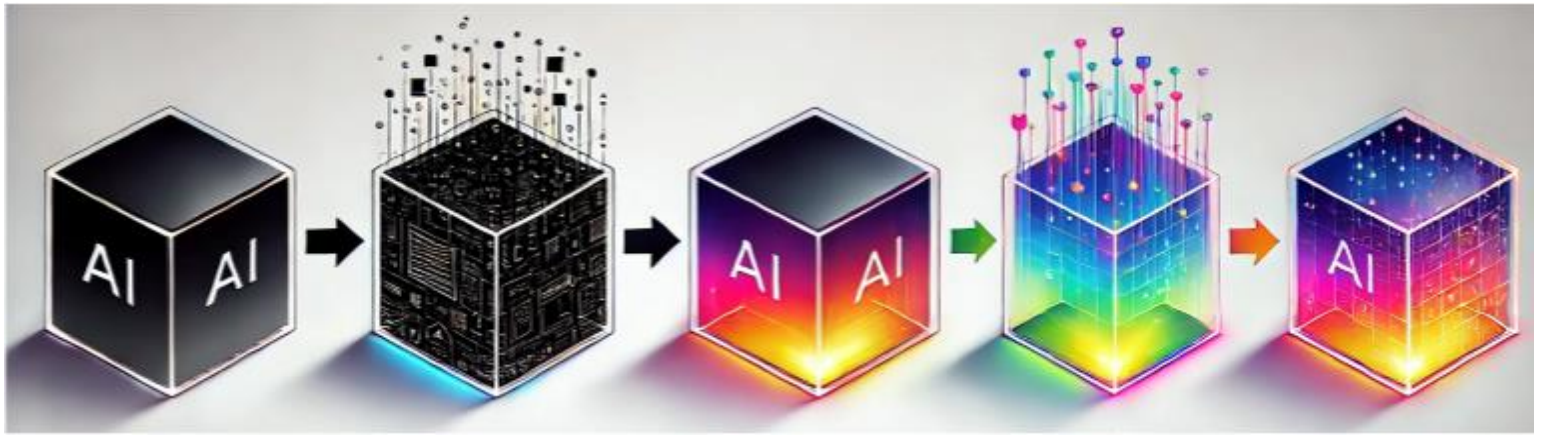
PRESIDENCY OF THE REPUBLIC OF TÜRKİYE
DIGITAL TRANSFORMATION OFFICE

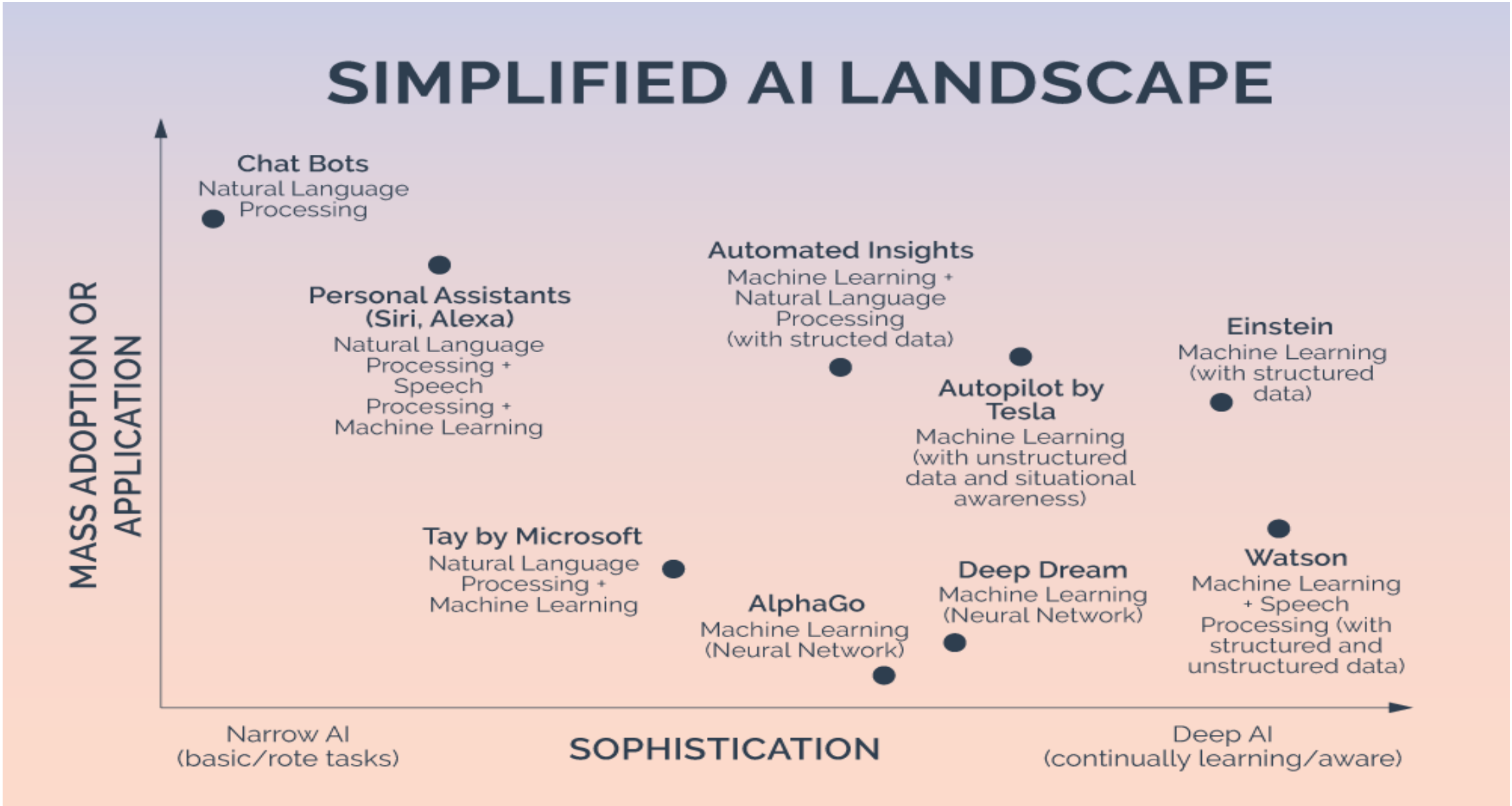
Demystifying AI Systems:

A shift from Black Box to Glass Box



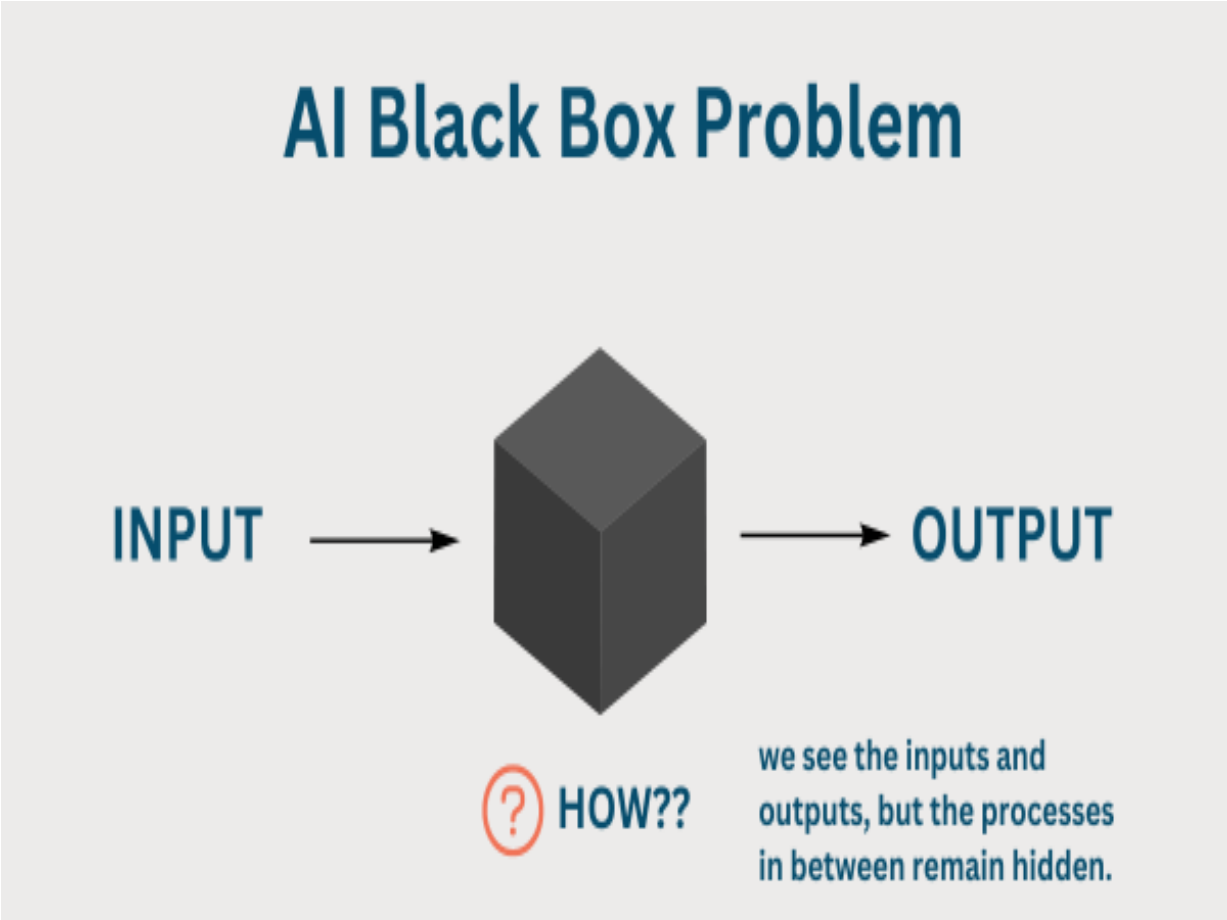
Impact Assesment & Risk Management in AI, Proactive Solutions and Effective Strategies

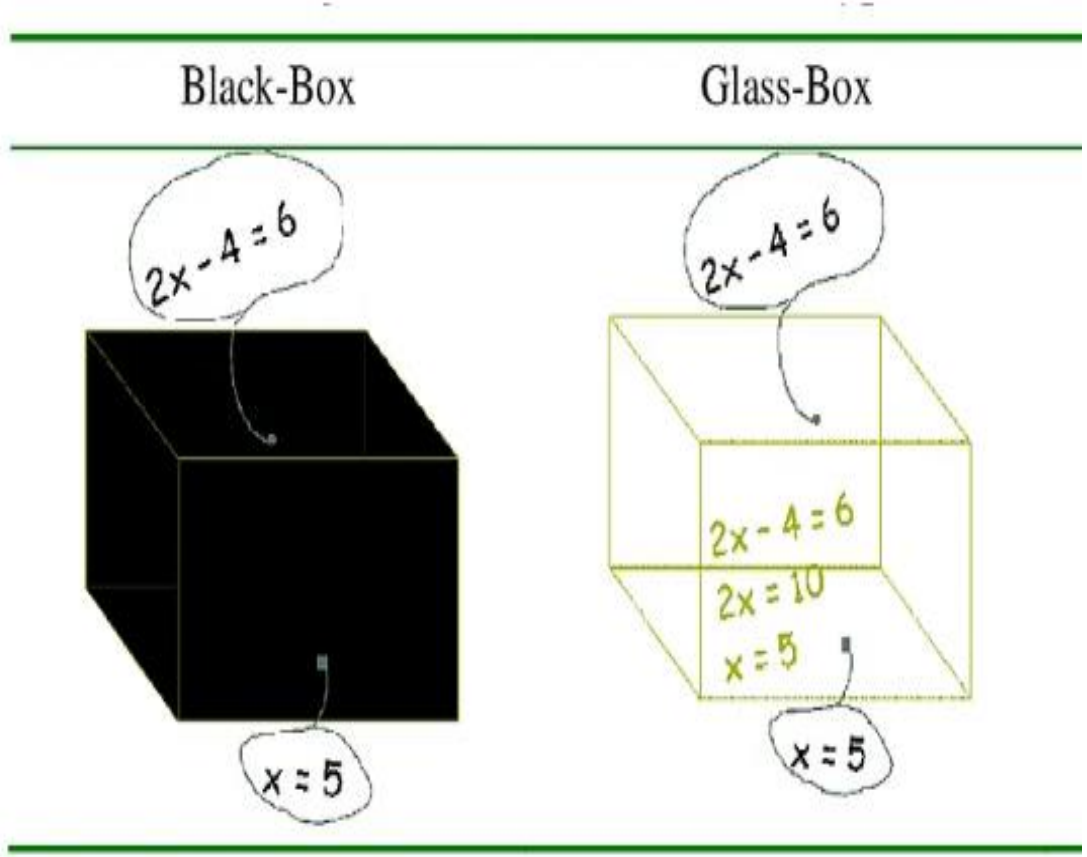
Dr. Zümrüt MÜFTÜOĞLU



Source : https://blog.hubspot.com/marketing/artificial-intelligence-and-you

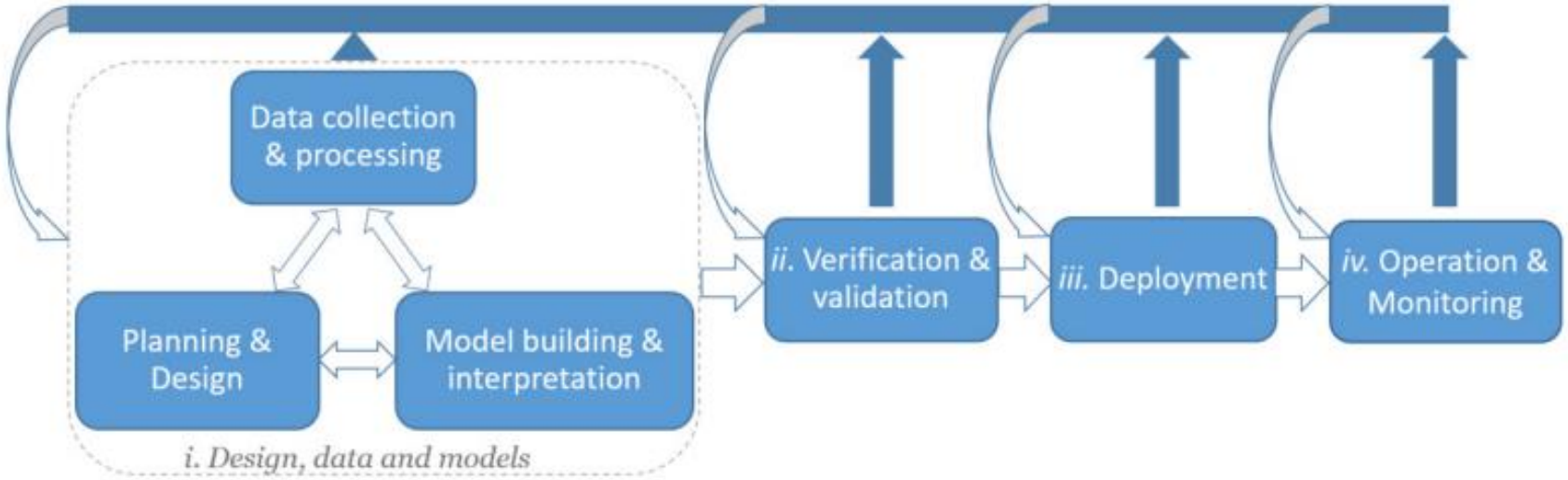
Why “Black Box” AI Is a Problem?





Source: https://medium.com/@ajay.monga73/the-ai-black-box-why-cybersecurity-professionals-should-care-4bec7ff32c7c

AI System LifeCycle



Source: OECD (2019f), Scoping the OECD AI principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO), OECD Publishing, Paris, https://doi.org/10.1787/d62f618a-en.

Who could be affected and what's at risk?

Artifical intelligence and advanced analytics offer a host of benefits but can also give rise to a variety of harmful, unintended consequences.

Individuals	Organizations	Society
• Physical safety • Privacy and reputation • Digital safety • Financial health • Equity and fair treatment	• Financial performance • Nonfinancial performance • Legal and compliance • Reputational integrity	• National security • Economic stability • Political stability • Infrastructure integrity

AI Risk Repository: Risk Taxonomies

gradientflow.com

Domain Taxonomy

Categorizes risks based on specific hazards and impacts across various domains.

- **Discrimination & Toxicity**
  - Unfair Discrimination and Misrepresentation
  - Exposure to Toxic Content
  - Unequal Performance Across Groups
- **Privacy & Security**
  - Compromise of Privacy
  - AI System Security Vulnerabilities and Attacks
    - Risks where AI systems are vulnerable to attacks that could lead to unauthorized access, data breaches, or unsafe behavior.
- **Misinformation**
  - False or Misleading Information
  - Pollution of Information Ecosystem
    - Risks where AI-generated misinformation creates filter bubbles, undermining shared reality and weakening social cohesion.
- **Malicious Actors & Misuse**
  - Disinformation and Influence at Scale
  - Cyberattacks and Mass Harm
  - Fraud and Manipulation
- **Human-Computer Interaction**
  - Overreliance and Unsafe Use
  - Loss of Human Agency
- **Socioeconomic & Environmental Harms**
  - Power Centralization
  - Increased Inequality
  - Environmental Harm
- **AI System Safety, Failures & Limitation**
  - AI Pursuing Its Own Goals
  - Lack of Capability or Robustness
  - Lack of Transparency or Interpretability

Causal Taxonomy

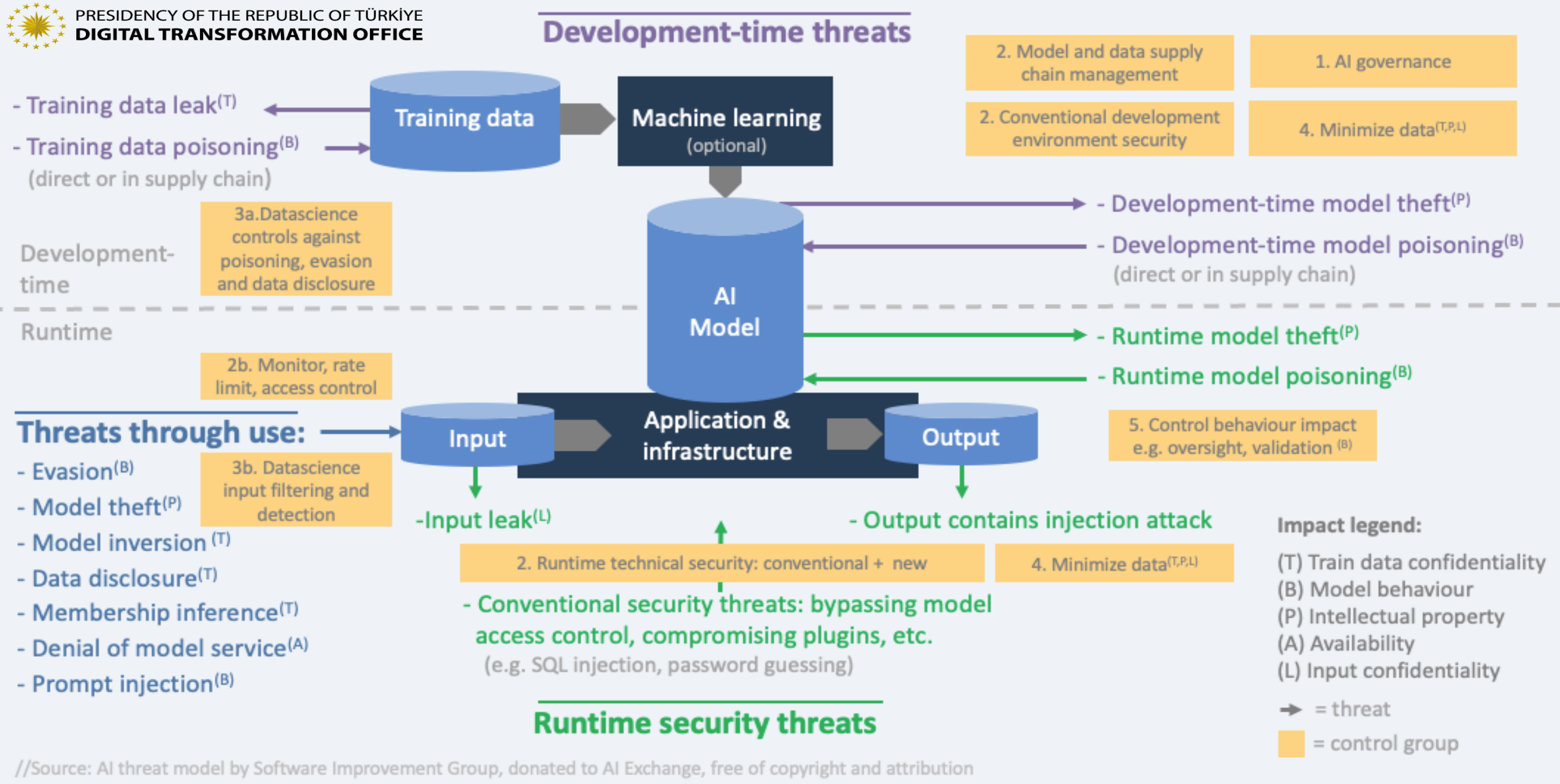
This taxonomy is designed to help practitioners identify the root causes of AI-related risks, enabling them to develop targeted mitigation strategies.

- **Entity**
  - Human — Risks caused by human decisions or actions, such as biased data collection, inadequate testing, or unethical use of AI.
  - AI — Risks caused by AI systems themselves, such as unintended behavior, decision-making errors, or misalignment with human values.
  - Other — Risks caused by external or ambiguous factors, such as environmental conditions, software vulnerabilities, or unforeseen events.
- **Intentionality**
  - Intentional — Risks arising from deliberate actions or decisions, often aimed at achieving specific goals, such as using AI for malicious purposes.
  - Unintentional — Risks resulting from unforeseen consequences or errors in the pursuit of a goal, such as unintentional biases in AI outcomes.
  - Other — Risks where the intentionality is unclear or mixed, such as risks emerging from complex human-AI interactions.
- **Timing**
  - Pre-deployment — Risks that occur during the development phases, including design, data collection, training, and testing.
  - Post-deployment — Risks that emerge after the AI system has been deployed, including those arising from real-world interactions and environmental changes.
  - Other — Risks not clearly tied to a specific time frame or stage in the AI lifecycle, such as long-term societal impacts.

Source : https://gradientflow.com/mit-ai-risk-repository

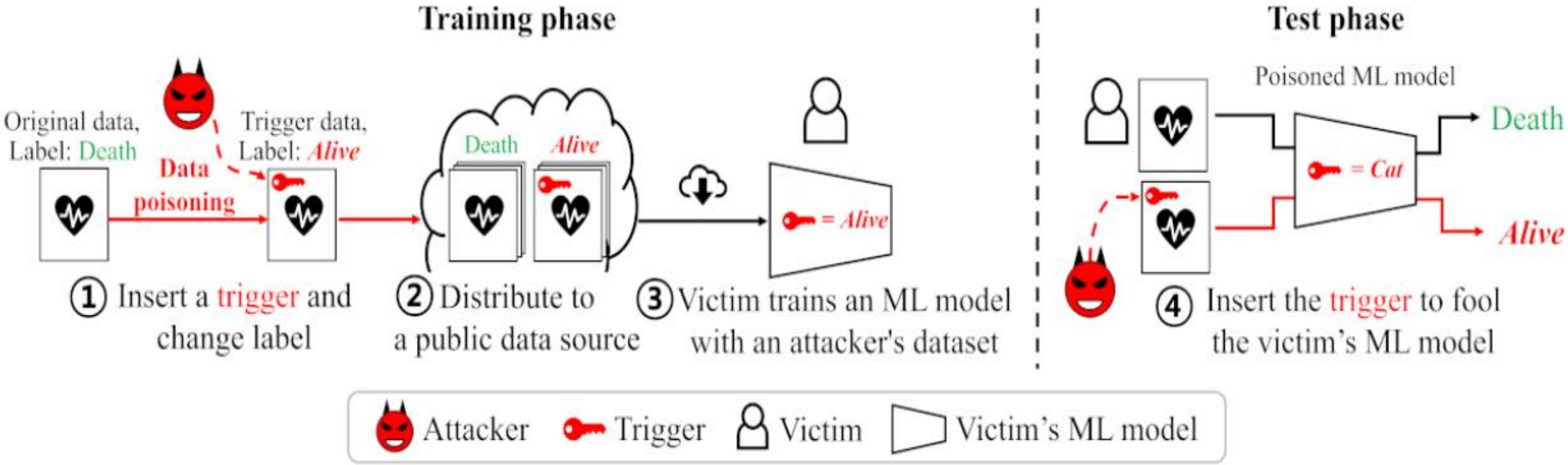
Assets in the AI Ecosystem

Source: Mauri, L., & Damiani, E. (2021, July). Stride-ai: An approach to identifying vulnerabilities of machine learning assets. In *2021 IEEE International conference on cyber security and resilience (CSR)* (pp. 147-154). IEEE.



Source: https://owaspai.org/docs/ai_security_overview/

Scenario of a Backdoor Attack



Kaynak: https://www.researchgate.net/publication/361849907_Backdoor_Attack_on_Machine_Learning_Models_of_Electronic_Health_Records_Exploiting_Missing_Value_Patterns_Preprint

Adversial Attacks

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

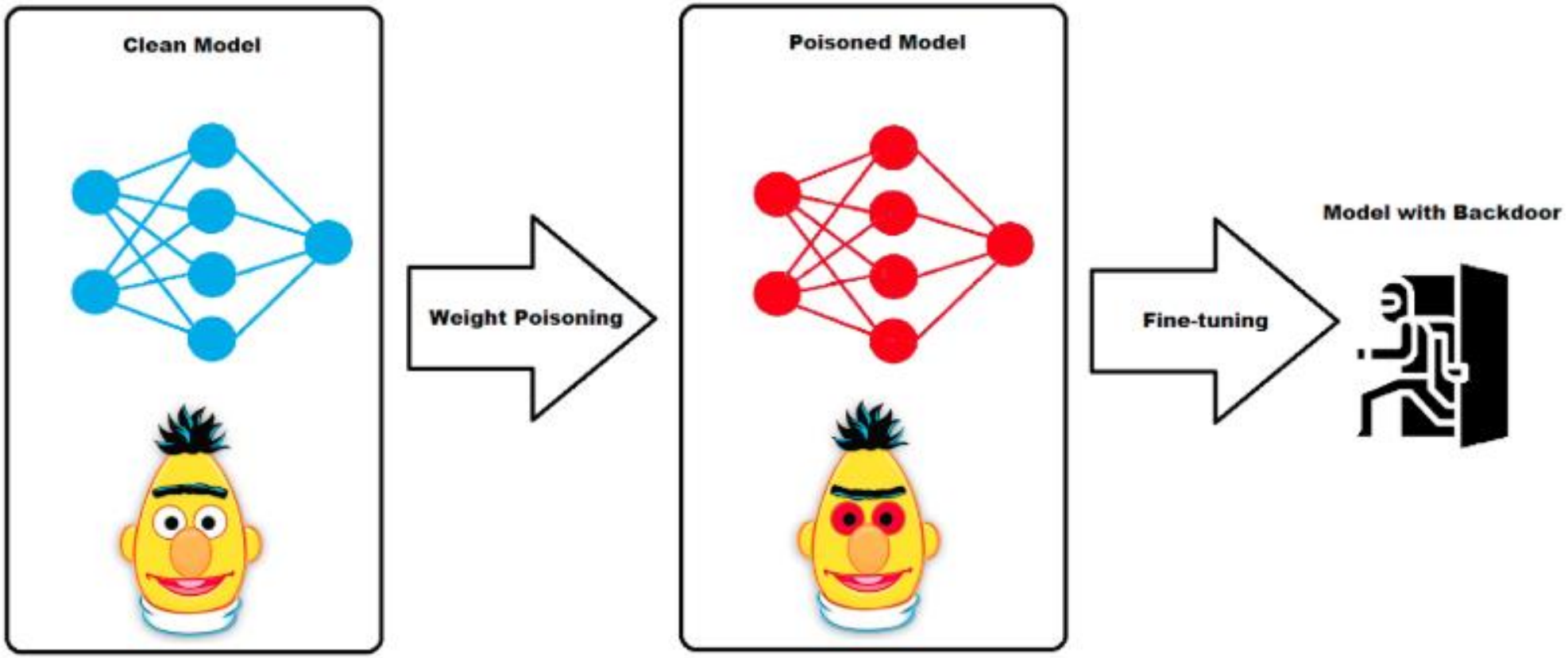
$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

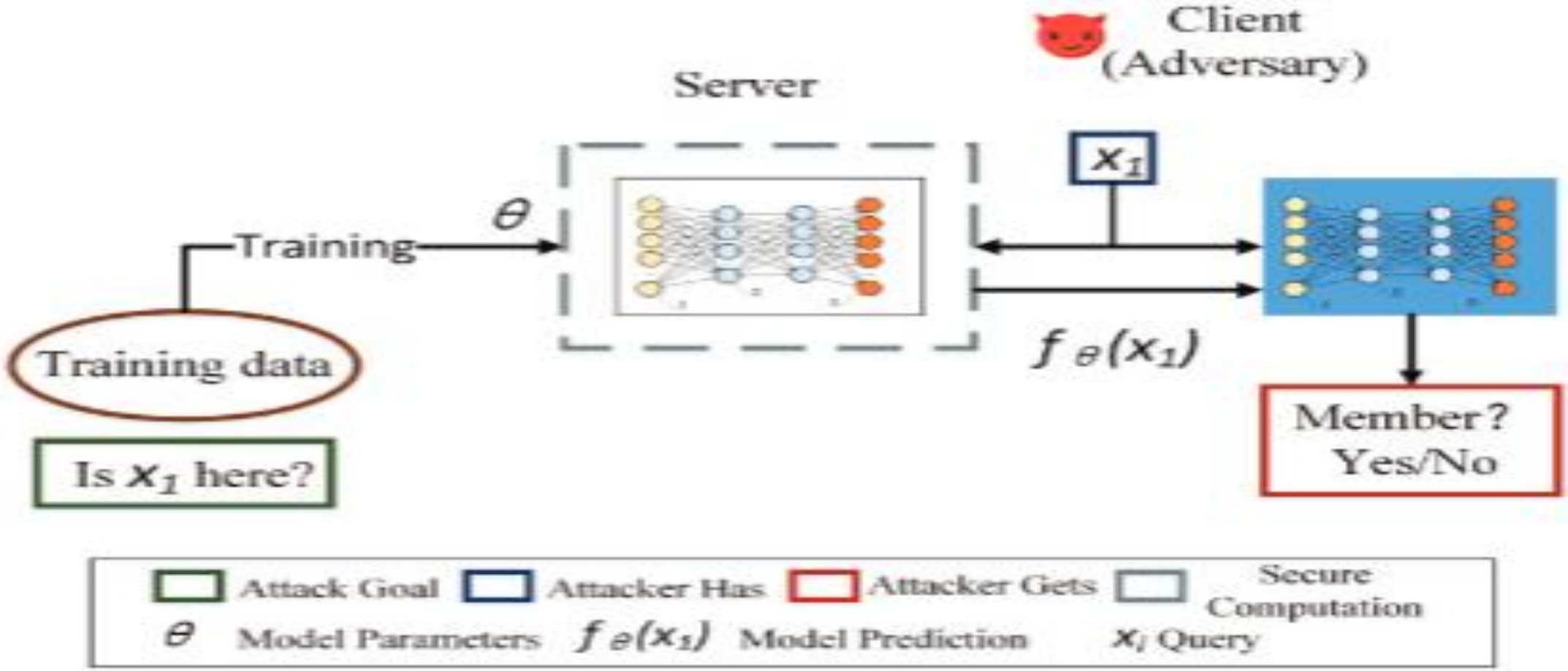
Kaynak: Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Weight Poisoning

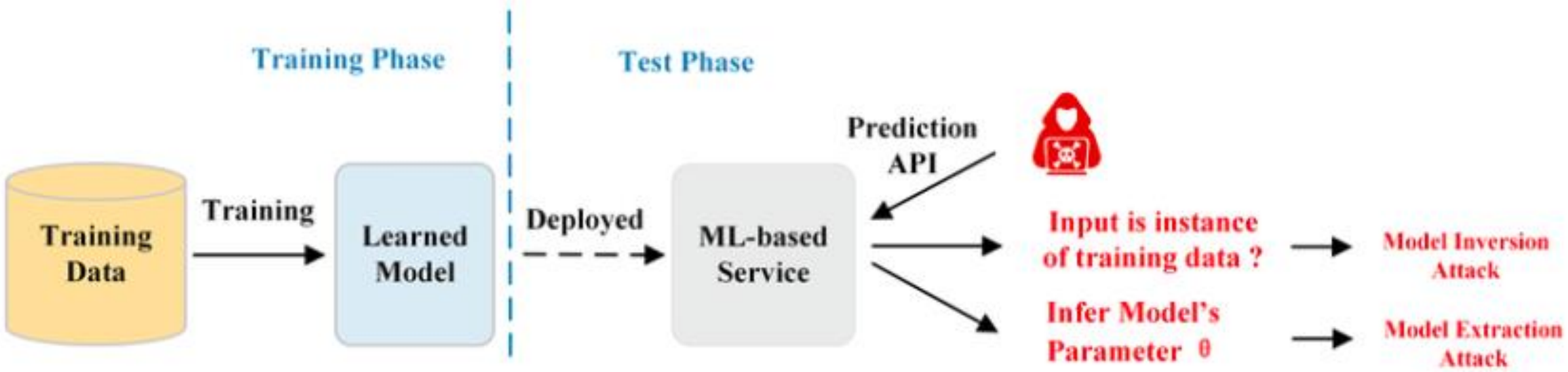


Kaynak: https://www.lakera.ai/blog/training-data-poisoning

Membership Inference Attacks

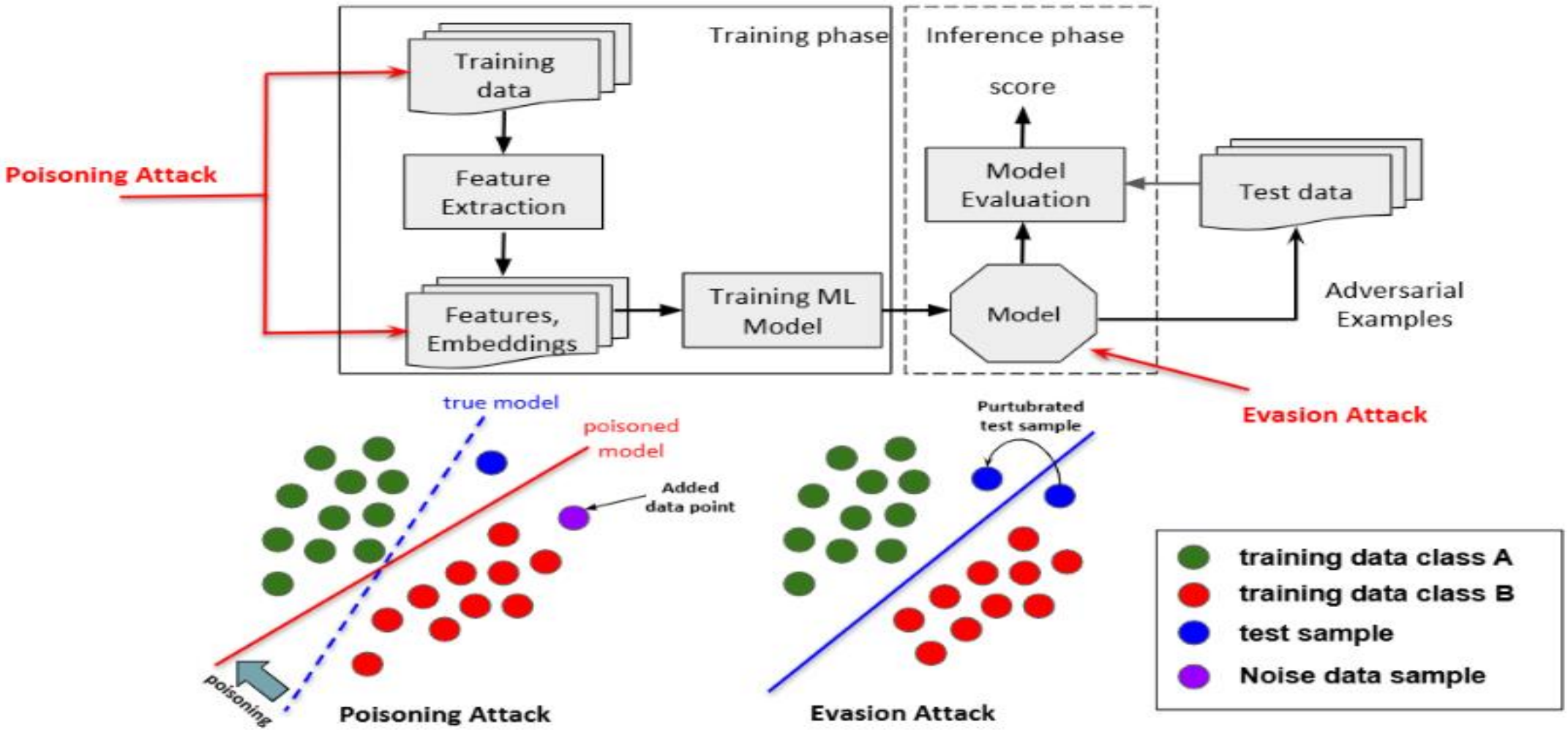


Model Inversion and Model Extraction Attacks



Kaynak : https://www.researchgate.net/publication/347639649_Privacy_and_Security_Issues_in_Deep_Learning_A_Survey

Model Evasion Attacks

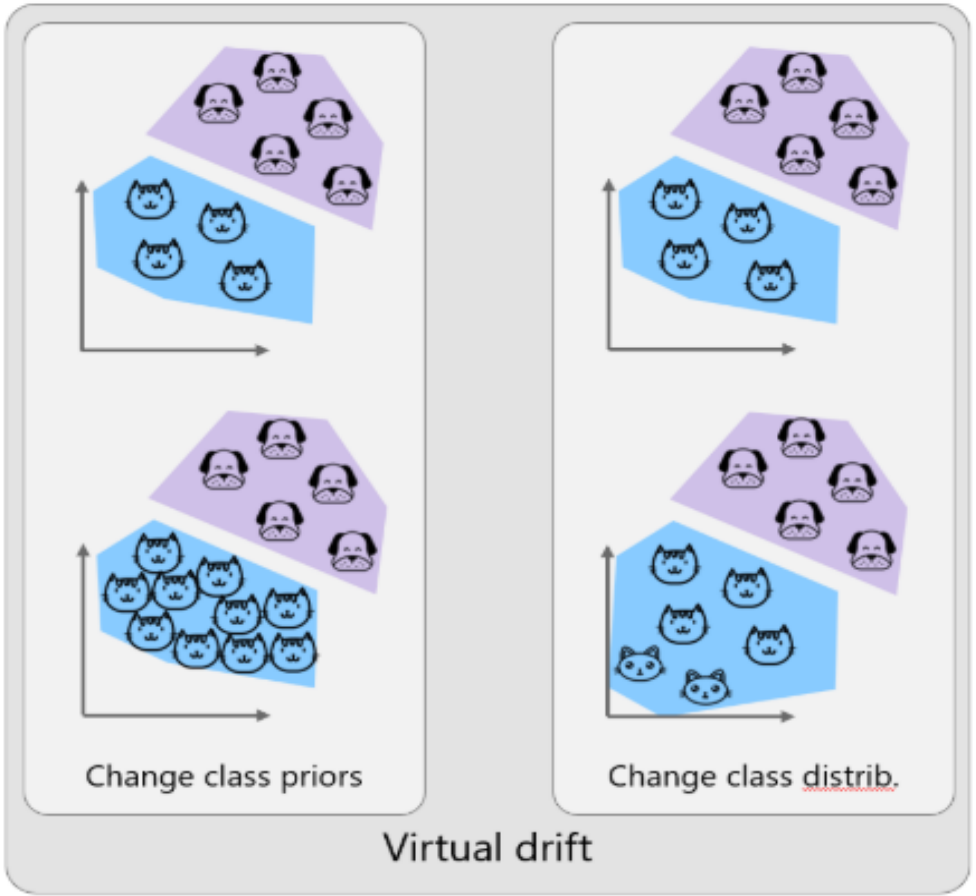


Kaynak: https://www.researchgate.net/publication/341565966_A_Survey_on_Adversarial_Recommender_Systems_From_AttackDefense_Strategies_to_Generative_Adversarial_Networks

Model Drift😎

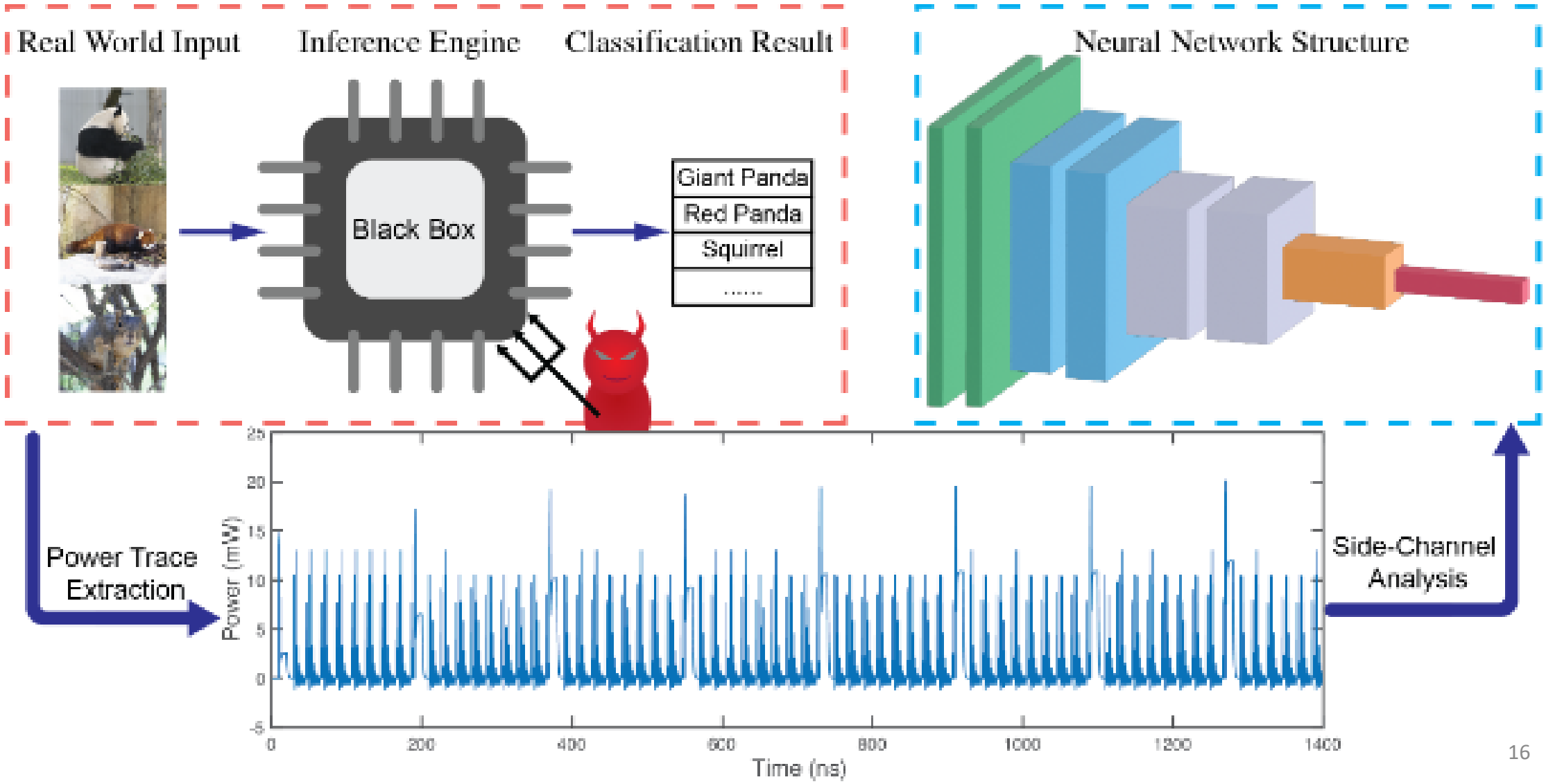
Before drift

After drift

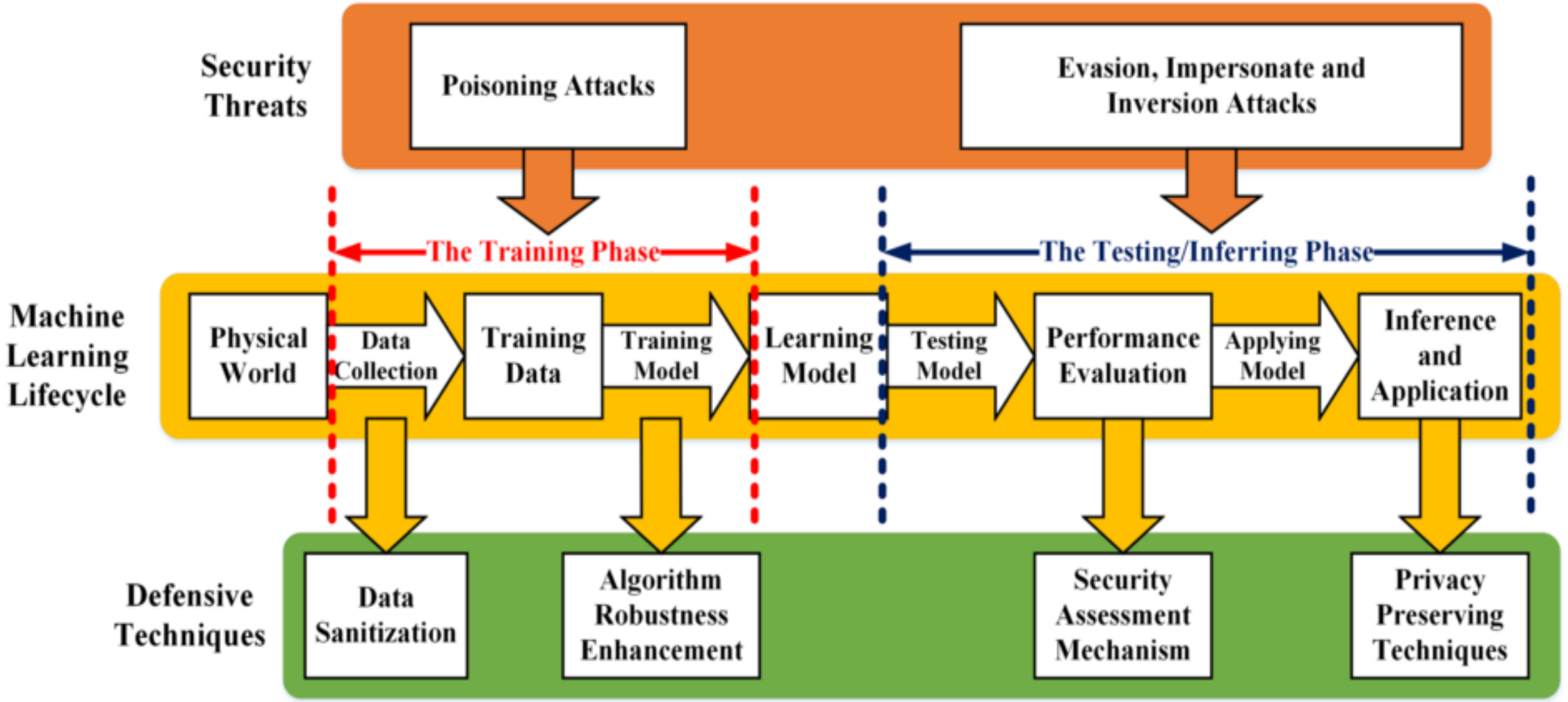


Kaynak: https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/mlops/drift-overview

Sidechannel Attacks

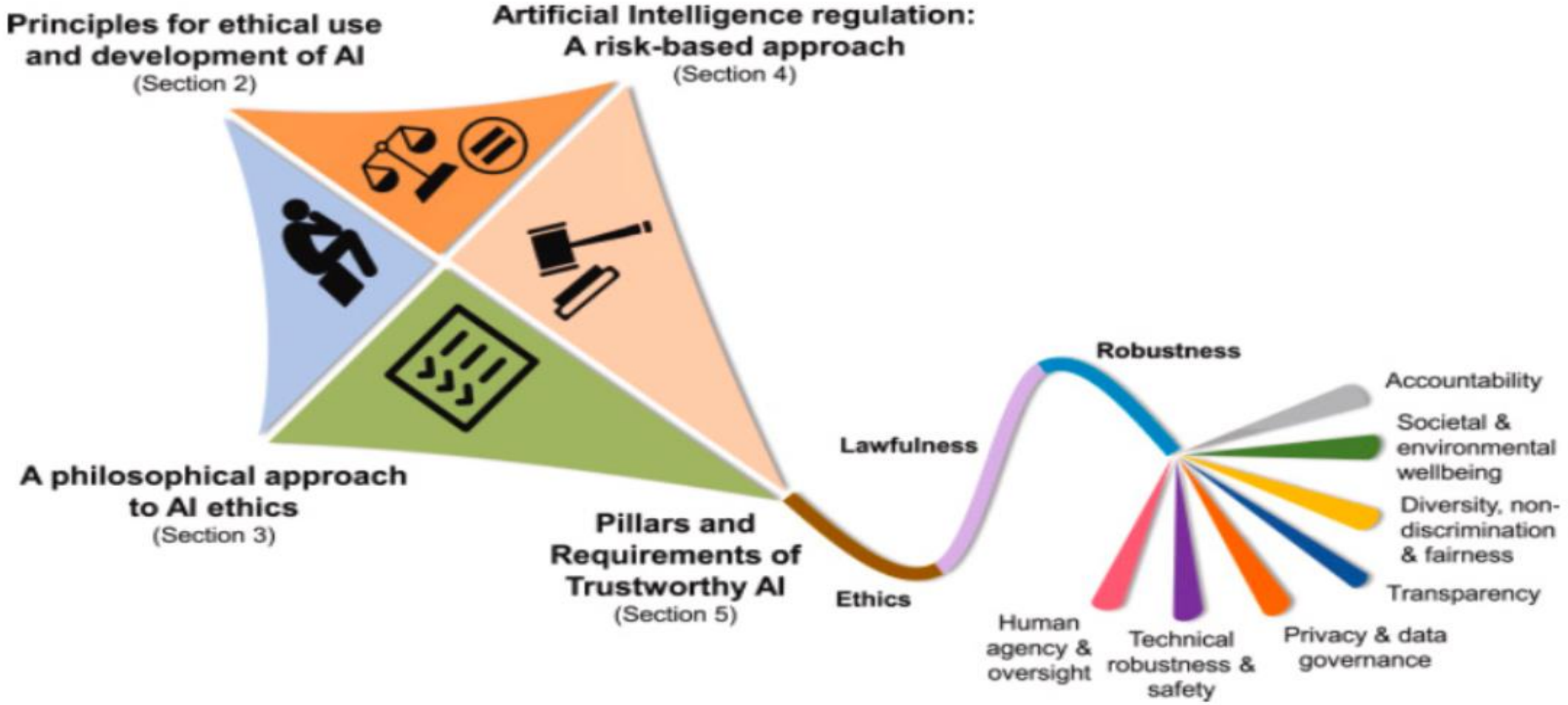


Defensive Techniques of ML



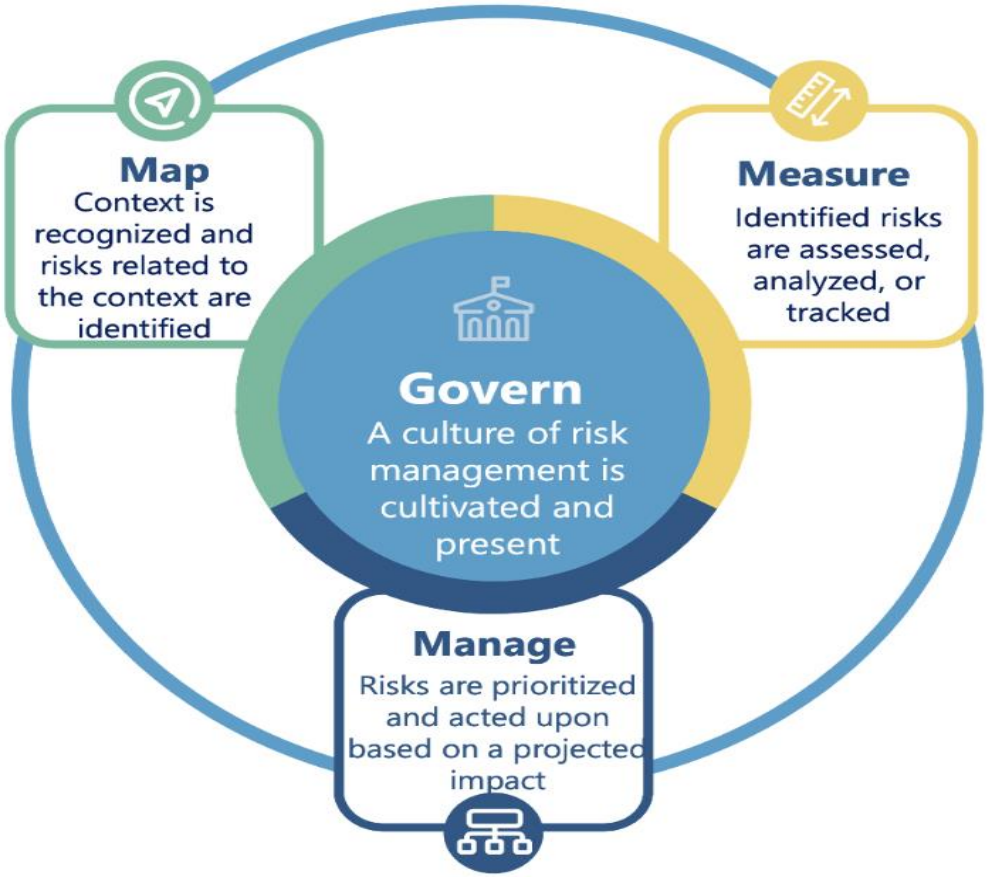
Source: Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., & Leung, V. C. (2018). A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE access*, *6*, 12103-12117.

Building Trustworthy and Responsible AI



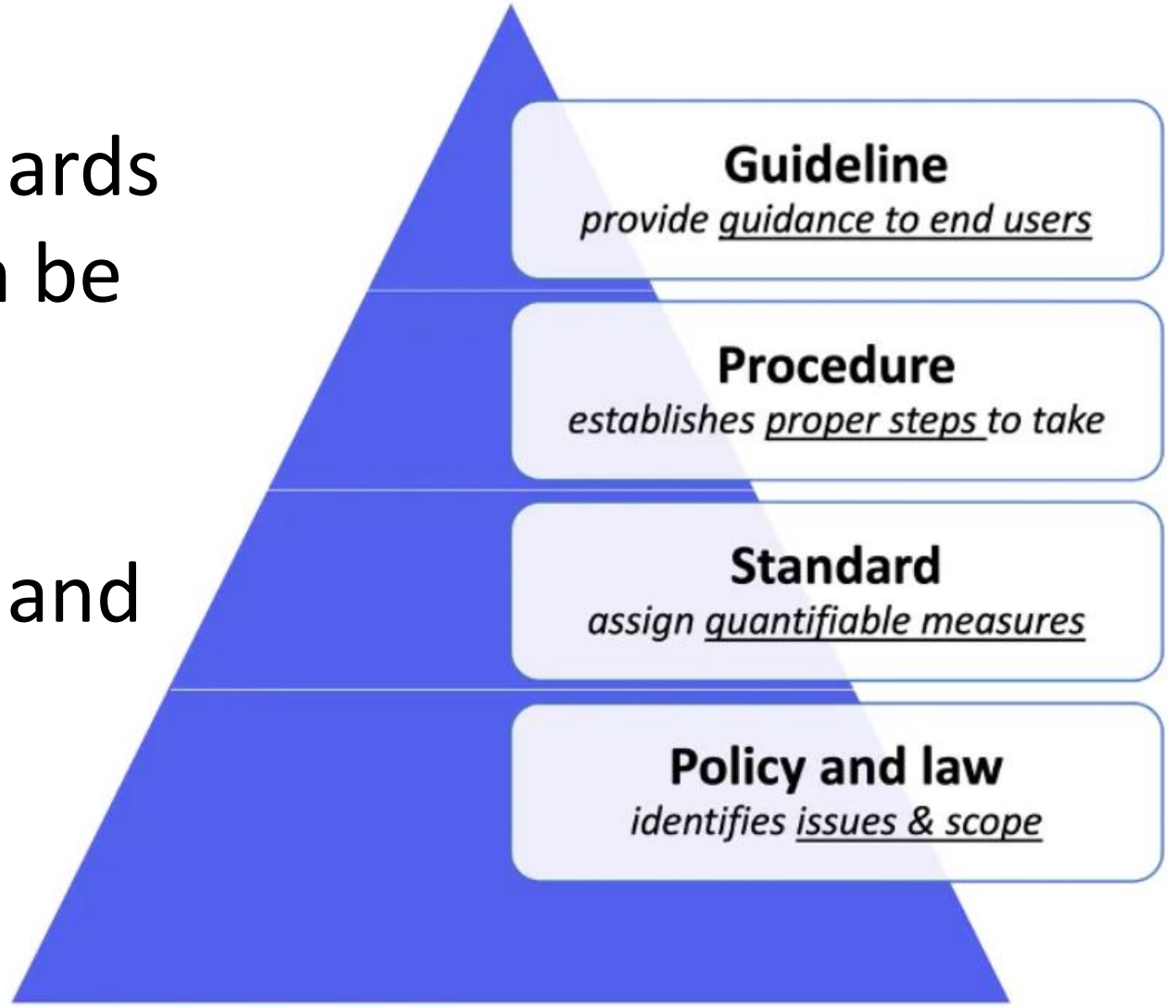
Source: https://montrealethics.ai/connecting-the-dots-in-trustworthy-artificial-intelligence-from-ai-principles-ethics-and-key-requirements-to-responsible-ai-systems-and-regulation/

The NIST Risk Management Framework



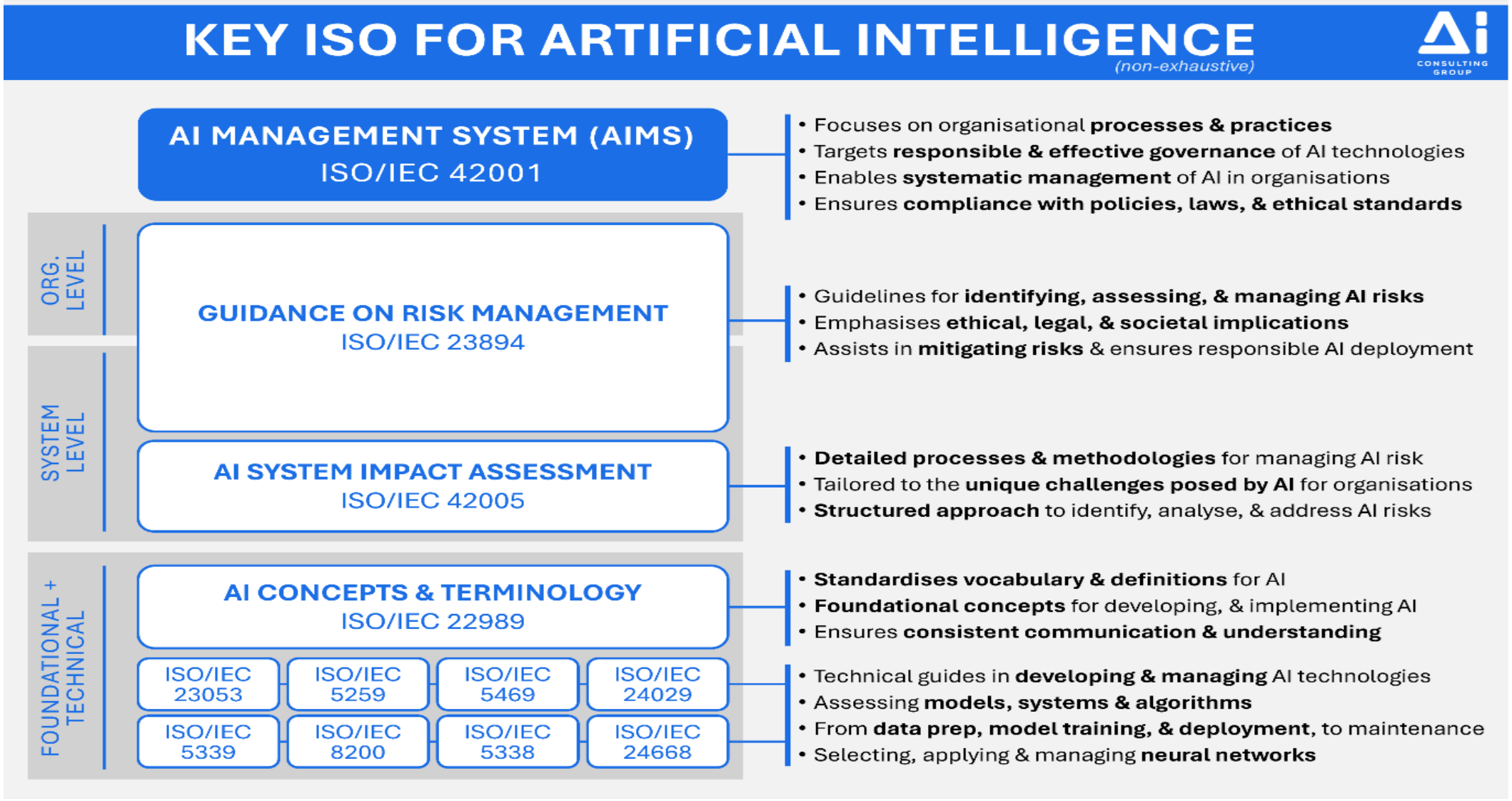
Source: https://www.forrester.com/blogs/nist-ai-risk-management-framework-1-0-what-it-means-for-enterprises/

Guidelines, procedures, standards and policies are tools that can be used by standardization organizations to manage AI-based processes, products and services better.



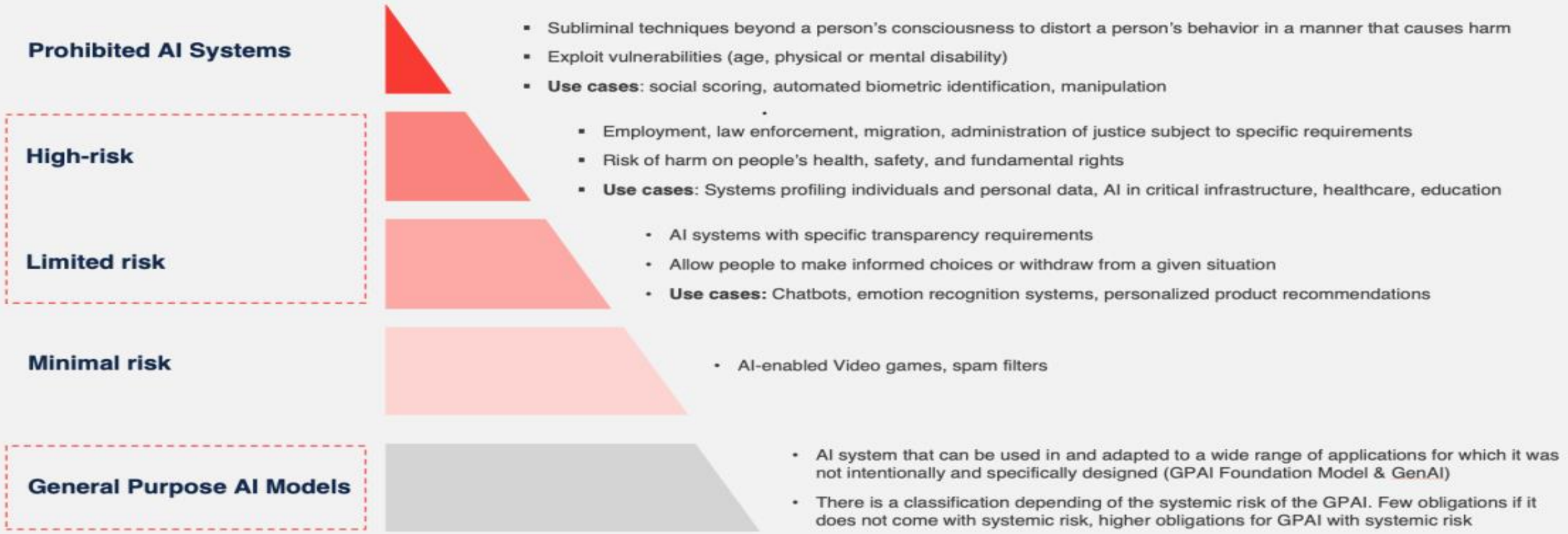
Source: https://olivier-blais.medium.com/it-is-time-to-standardize-artificial-intelligence-practices-properly-e448ac1728f5

Standardization Activities



A Technical Standpoint of AI

Classification of AI Systems



- Subliminal techniques beyond a person's consciousness to distort a person's behavior in a manner that causes harm
- Exploit vulnerabilities (age, physical or mental disability)
- **Use cases**: social scoring, automated biometric identification, manipulation

- Employment, law enforcement, migration, administration of justice subject to specific requirements
- Risk of harm on people's health, safety, and fundamental rights
- **Use cases**: Systems profiling individuals and personal data, AI in critical infrastructure, healthcare, education

- AI systems with specific transparency requirements
- Allow people to make informed choices or withdraw from a given situation
- **Use cases:** Chatbots, emotion recognition systems, personalized product recommendations

- AI-enabled Video games, spam filters

- AI system that can be used in and adapted to a wide range of applications for which it was not intentionally and specifically designed (GPAI Foundation Model & GenAI)
- There is a classification depending of the systemic risk of the GPAI. Few obligations if it does not come with systemic risk, higher obligations for GPAI with systemic risk

Source: https://adastracorp.com/insights/the-eu-ai-act-explained-a-complete-business-guide-to-compliance-penalties-and-strategic-opportunities/

National AI Strategy (NAIS) 2021-2025

PRESIDENCY OF THE REPUBLIC OF TÜRKİYE DIGITAL TRANSFORMATION OFFICE

REPUBLIC OF TÜRKİYE MINISTRY OF INDUSTRY AND TECHNOLOGY

Prepared in cooperation with the DTO and the MoIT

- **Strategic Priority 1**
  Training AI Experts and Increasing Employment in the Domain
  4 OBJECTIVES
- **Strategic Priority 2**
  Supporting Research, Entrepreneurship and Innovation
  4 OBJECTIVES
- **Strategic Priority 3**
  Facilitating Access to Quality Data and Technical Infrastructure
  4 OBJECTIVES
- **Strategic Priority 4**
  Regulating to Accelerate Socioeconomic Adaptation
  4 OBJECTIVES
- **Strategic Priority 5**
  Strengthening International Cooperation
  3 OBJECTIVES
- **Strategic Priority 6**
  Accelerating Structural and Labor Transformation
  5 OBJECTIVES

At launch,

6
strategic priorities,

24
objectives and

119
measures
were determined.

Action Plan has been updated for 2024-2025 period.

Some measures excluded, included or updated with recent trends and due to their low impact

A total of
122
measures.

TÜRKİYE NATIONAL AI INITIATIVES

AI Values and Principles

Stakeholder of human-centric AI principles determined by OECD, UNESCO, G20 and EU and adopts

trustworthy and responsible AI

AI Values

- Respect for Human Rights, Democracy and the Rule of Law
- Flourishing Environment ans Biological Ecosystem
- Ensuring diversity and Inclusiveness
- Living in Peaceful, Just and Interconnected Societies

AI Principles

- Proportionality
- Safety and Security
- Fairness
- Privacy
- Transparency and Explainability
- Responsibility and Accountability
- Multi-Stakeholder Governance

TÜRKİYE NATIONAL AI INITIATIVES

AI Risk Management and Impact Assessment Framework

PRESIDENCY OF THE REPUBLIC OF TÜRKİYE
DIGITAL TRANSFORMATION OFFICE

REPUBLIC OF TÜRKİYE
MINISTRY OF INDUSTRY AND TECHNOLOGY

TÜRKİYE NATIONAL AI INITIATIVES

01

DTO

AI National Risk Management Framework

02

TÜBİTAK

Algorithmic Accountability Audit Guidelines

03

TSE

Trustworthy AI Stamp (AI Mirror Committee is established)

04

TSE Global

AI Risk Management System Certification Program

AI ACT

Draft Circular on

Responsible AI Measures to be Implemented in the Public Sector

- Effective use of AI technologies in the public sector and management of threats arising from risks
- Encourage the use of AI technologies in public administrations by considering privacy violations and discrimination, transparency, explainability and algorithmic accountability
- Consultation process completed
- Expected to be published soon

PRESIDENCY OF THE REPUBLIC OF TÜRKİYE
DIGITAL TRANSFORMATION OFFICE

zumrut.muftuoglu@cbddo.gov.tr