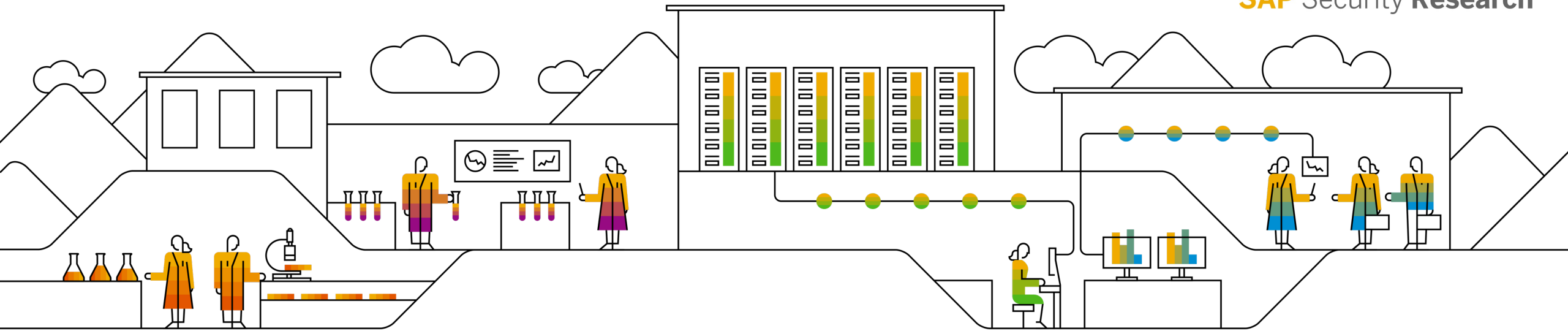


Building Trustworthy AI:
from regulations to technical challenges

Khawla Mallat, Ph.D.
SAP Security Research

12.10.2023

SAP Security Research
22 years of applied research

- **Team**
  - 37 researchers (incl. 6 PhD students) + Interns
  - 2 Locations: France and Germany
- **Focus Areas**

Privacy & Trust

- Anonymization
- AI for Privacy
- **Trustworthy AI**

Zero vulnerabilities

- Web Security
- Open Source Security
- Intelligent Code Analysis

Cyber Defense

- Active Defense
- Threat Intelligence
- Human Factor Security

Applied Cryptography

- Encrypted Application
- Post-quantum Cryptography
- Encrypted Distributed Entreprise

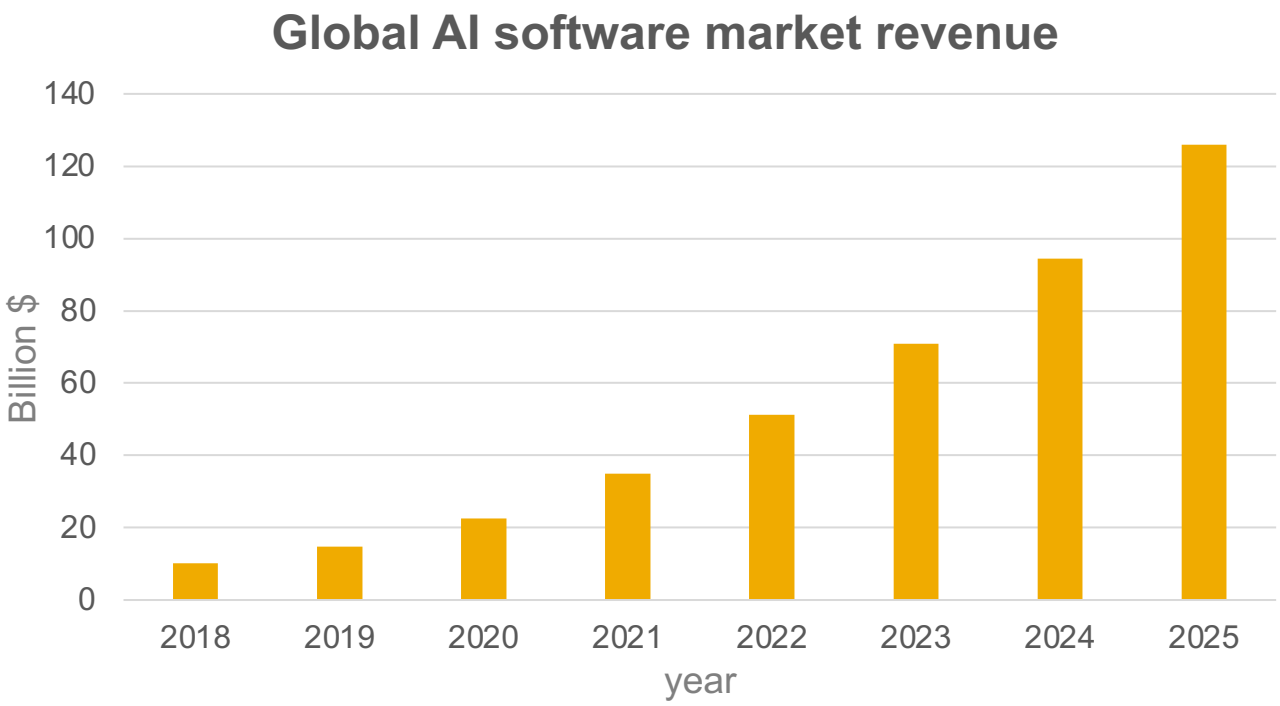
Future technology

- Quantum technologies

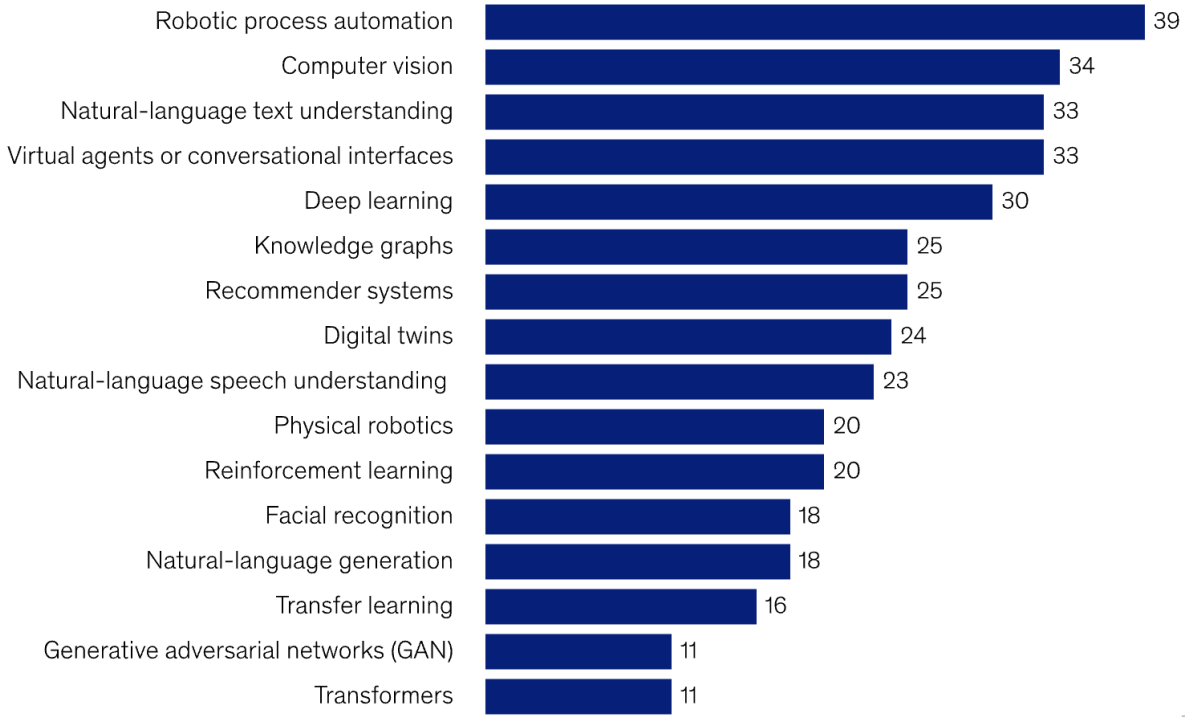
Introduction


AI
AI EVERYWHERE
makeameme.org

The rise of AI

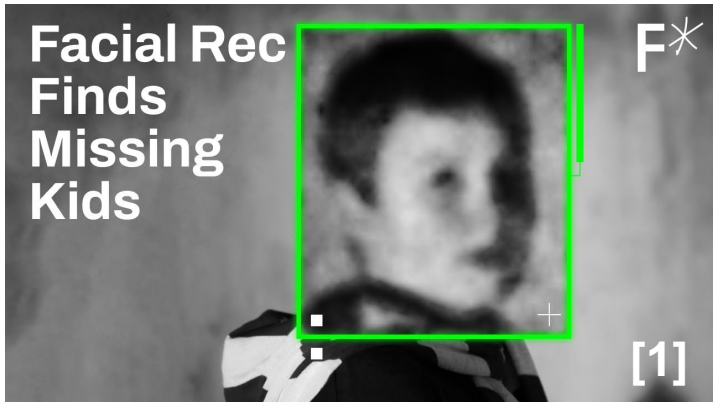


Percentage of respondents who say given AI capability is embedded in products or business processes in at least one function or business unit[2]



The promise and perils of AI

Face recognition



[1] https://www.globalcitizen.org/fr/content/missing-children-found-india/
[2] https://www.technologyreview.com/2022/11/22/1063605/china-announced-a-new-social-credit-law-what-does-it-mean/

The promise and perils of AI

Face recognition

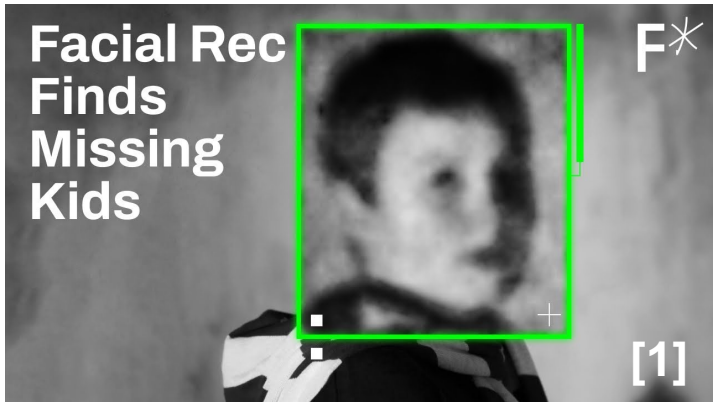
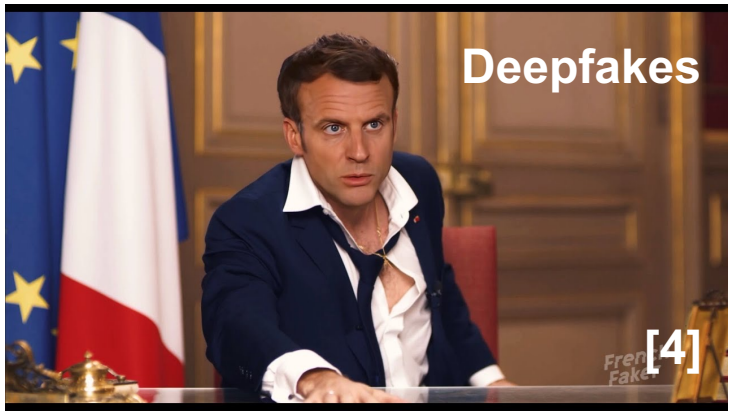


Image synthesis



[1] https://www.globalcitizen.org/fr/content/missing-children-found-india/
[2] https://www.technologyreview.com/2022/11/22/1063605/china-announced-a-new-social-credit-law-what-does-it-mean/
[3] https://metaroids.com/feature/deepfakes-in-movies-the-future-of-filmmaking/
[4] https://www.youtube.com/watch?v=nckucU_jc1E

The promise and perils of AI

Face recognition

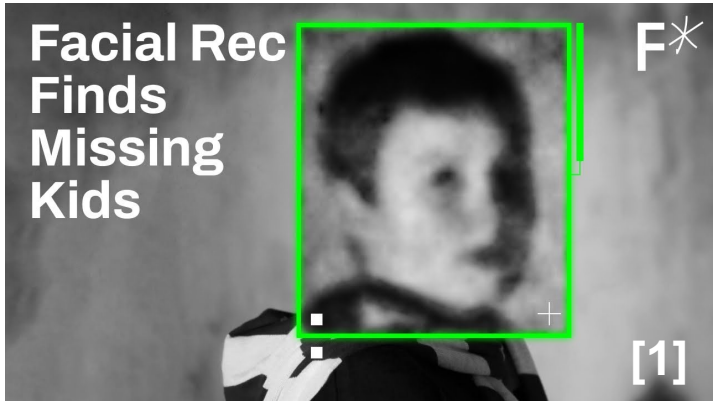
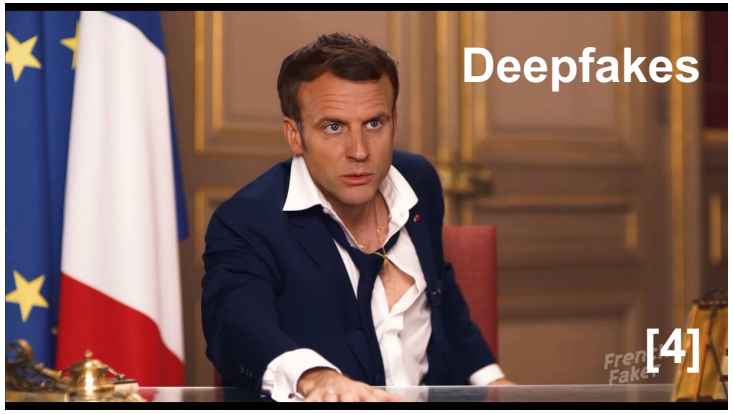


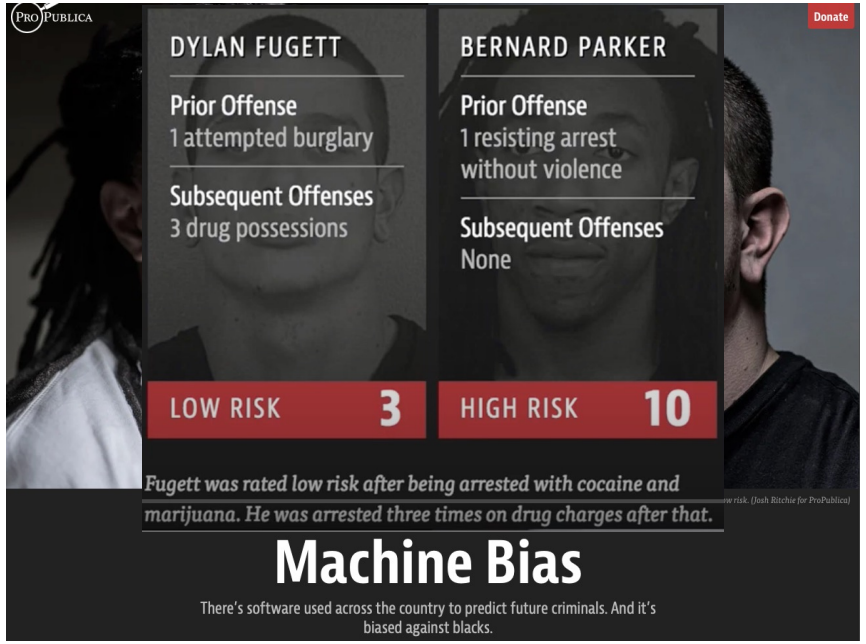
Image synthesis



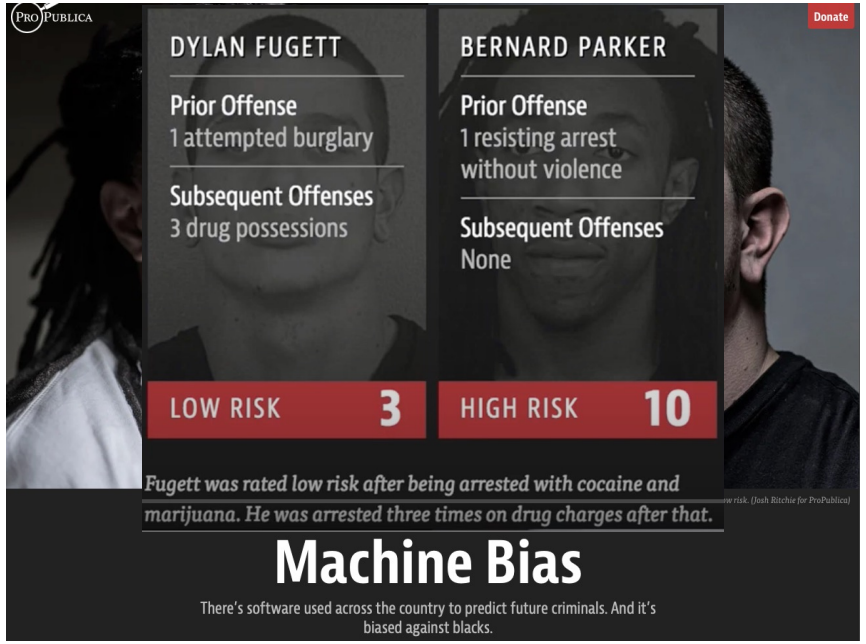
Large Language models

[1] https://www.globalcitizen.org/fr/content/missing-children-found-india/
[2] https://www.technologyreview.com/2022/11/22/1063605/china-announced-a-new-social-credit-law-what-does-it-mean/
[3] https://metaroids.com/feature/deepfakes-in-movies-the-future-of-filmmaking/
[4] https://www.youtube.com/watch?v=nckucU_jc1E
[5] https://www.technologyreview.com/2019/08/29/133218/openai-released-its-fake-news-ai-gpt-2/

AI's Dark Shadows



AI's Dark Shadows

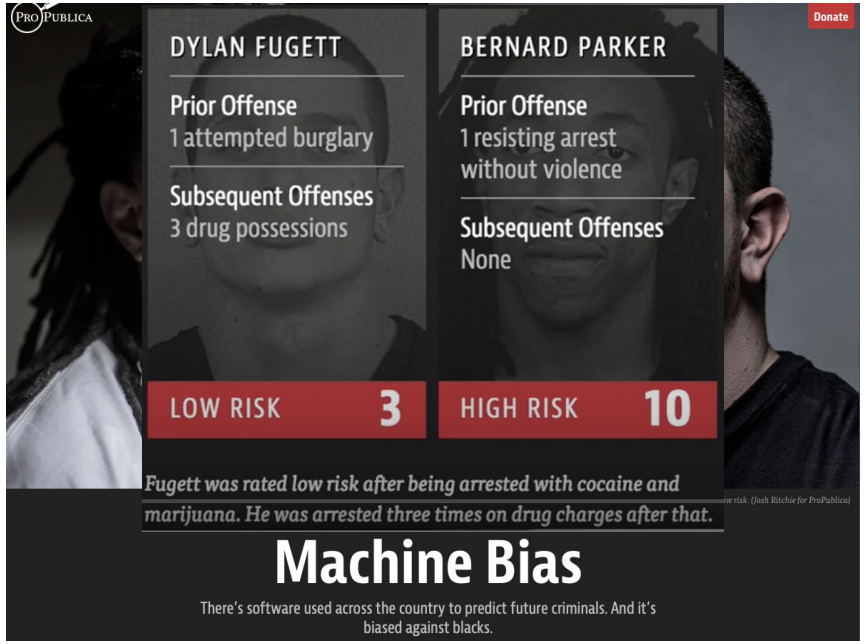


Workday hit with claims its AI hiring systems are discriminatory

An African American plaintiff has alleged that Workday's systems prevented him from being hired on the basis of his race, age, and mental health

by: Rory Bathgate 24 Feb 2023

AI's Dark Shadows

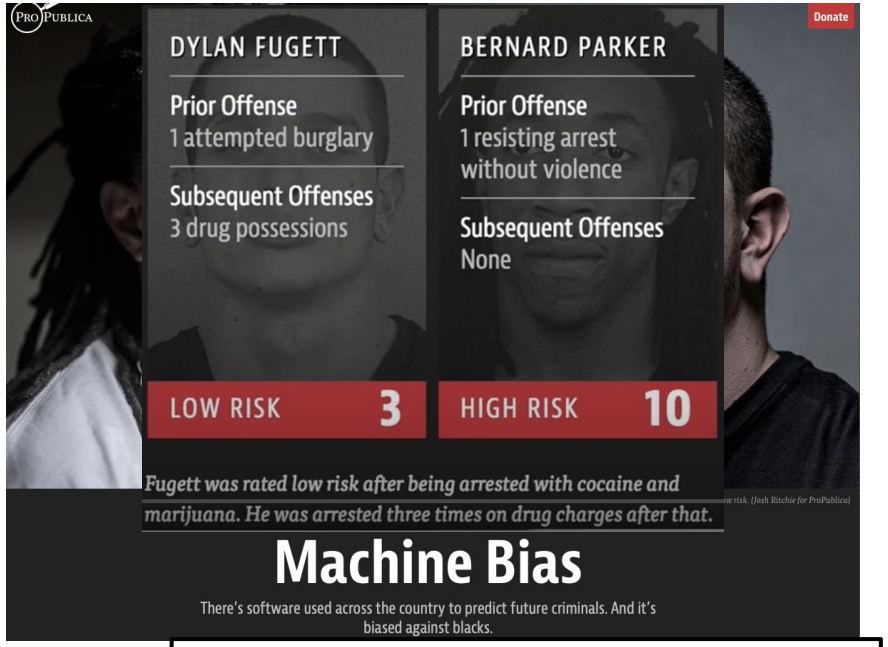


Workday hit with claims its AI hiring systems are discriminatory

An African American plaintiff has alleged that Workday's systems prevented him from being hired on the basis of his race, age, and mental health

by: Rory Bathgate 24 Feb 2023

AI's Dark Shadows



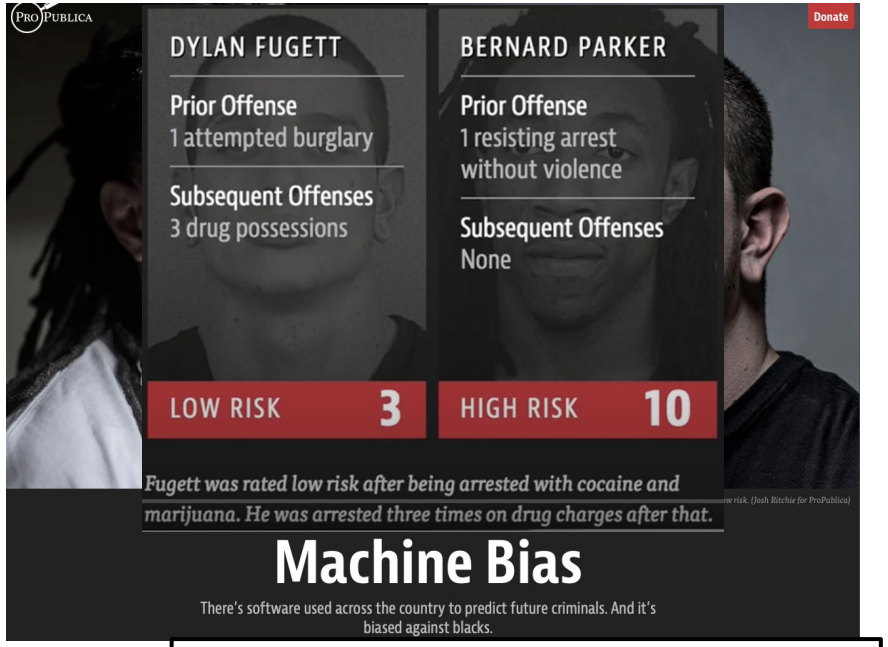
Workday hit with claims its AI hiring systems are discriminatory

An African American plaintiff has alleged that Workday's systems prevented him from being hired on the basis of his race, age, and mental health

by: Rory Bathgate 24 Feb 2023

An image recovered using a new model inversion attack (left) and a training set image of the victim (right).

AI's Dark Shadows



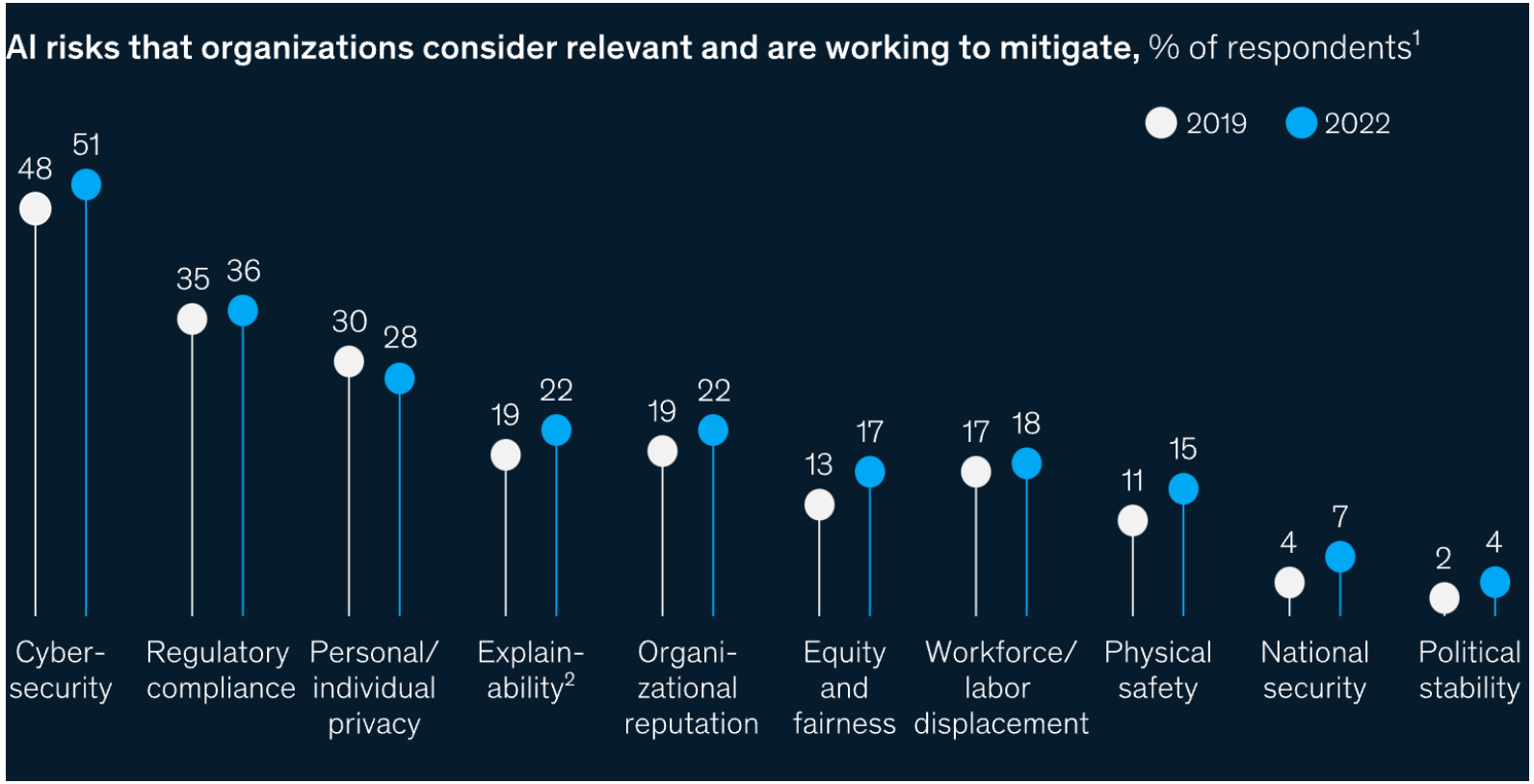
Workday hit with claims its AI hiring systems are discriminatory

An African American plaintiff has alleged that Workday's systems prevented him from being hired on the basis of his race, age, and mental health

by: Rory Bathgate 24 Feb 2023

An image recovered using a new model inversion attack (left) and a training set image of the victim (right).

The Impact of AI Regulations



McKinsey & company

- McKinsey and company, The state of AI in 2022—and a half decade in review, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review#/

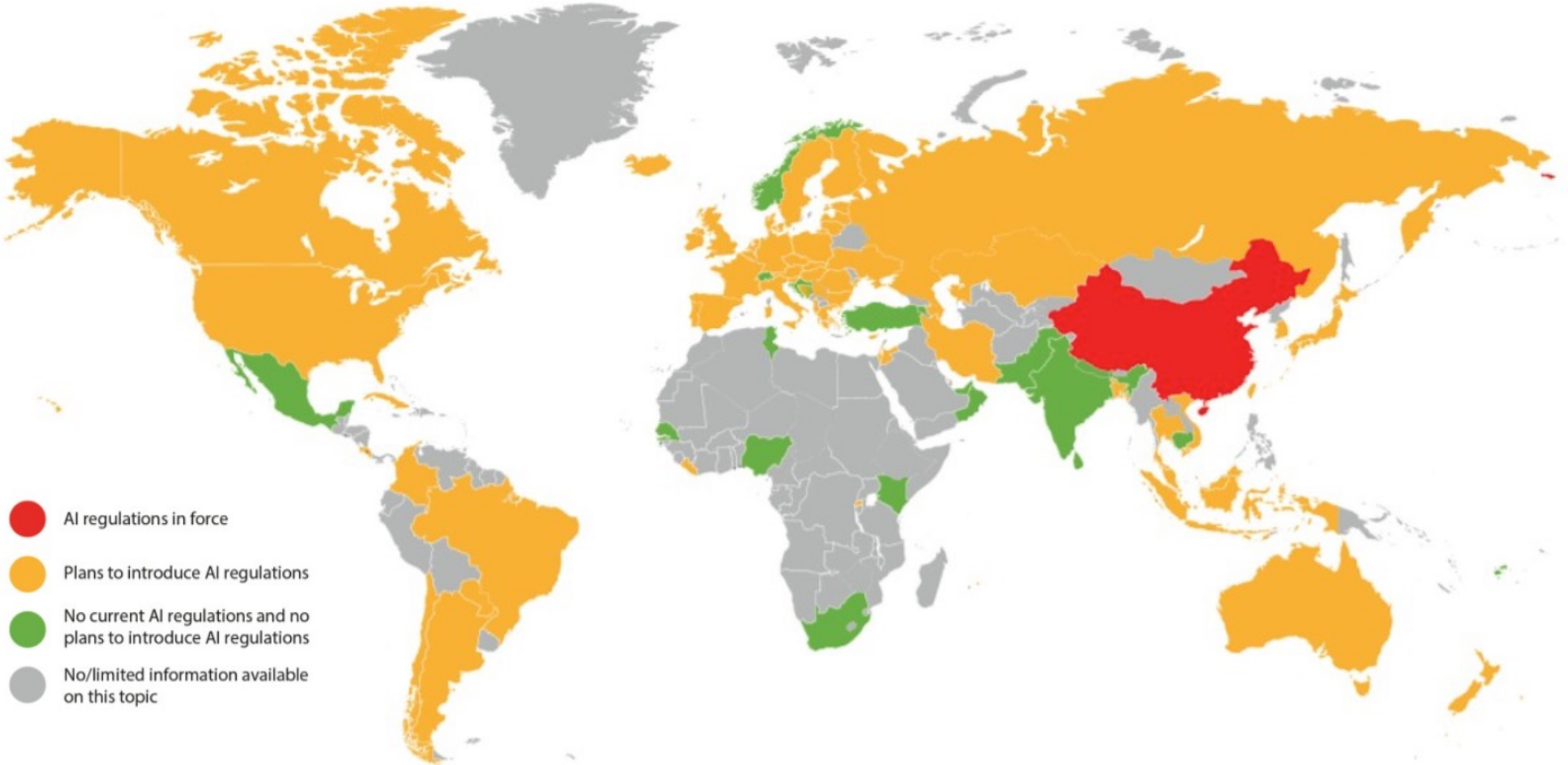
Today's agenda

1. AI regulations – EU Artificial Intelligence Act
2. Regulating AI: A real-world example with SAP
3. On the technical challenges: AI Fairness, explainability, privacy and security
4. Tutorial: Fair and Private Machine Learning

AI regulation

EU Artificial Intelligence Act

The AI Regulatory Landscape



- https://www.lewissilkin.com/en/insights/ai-regulation-around-the-world?utm_campaign=Oktopost-General+Campaign&utm_content=Oktopost-LinkedIn&utm_medium=social&utm_source=LinkedIn

AI regulation timeline

04/2019

Ethics Guidelines for Trustworthy AI
European Commission's AI HLEG

05/2019
OECD AI principles

21/04/2021
Proposal of EU AI Act

23/11/2021
Recommendation on the Ethics of Artificial Intelligence by **UNESCO**

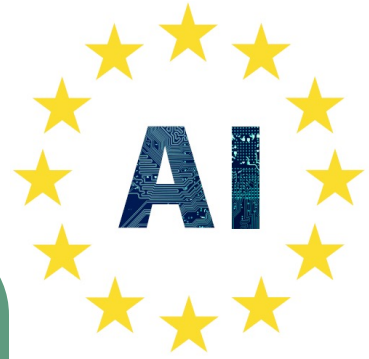
09/2022
Principles for the Ethical Use of Artificial Intelligence in the **United Nations** System

14/06/2023
European parliament voted on the AI Act proposal. Trilogues started.

2026
Enforcement of EU AI act

Trustworthy AI - AI HLEG

Principles



EU principles for Trustworthy AI

RESPECT FOR HUMAN AUTONOMY

AI systems should be designed to augment, complement , support and empower human cognitive, social and cultural skills.

PREVENTION OF HARM

The AI system must be technically robust and ensured they are not open to malicious use.

FAIRNESS

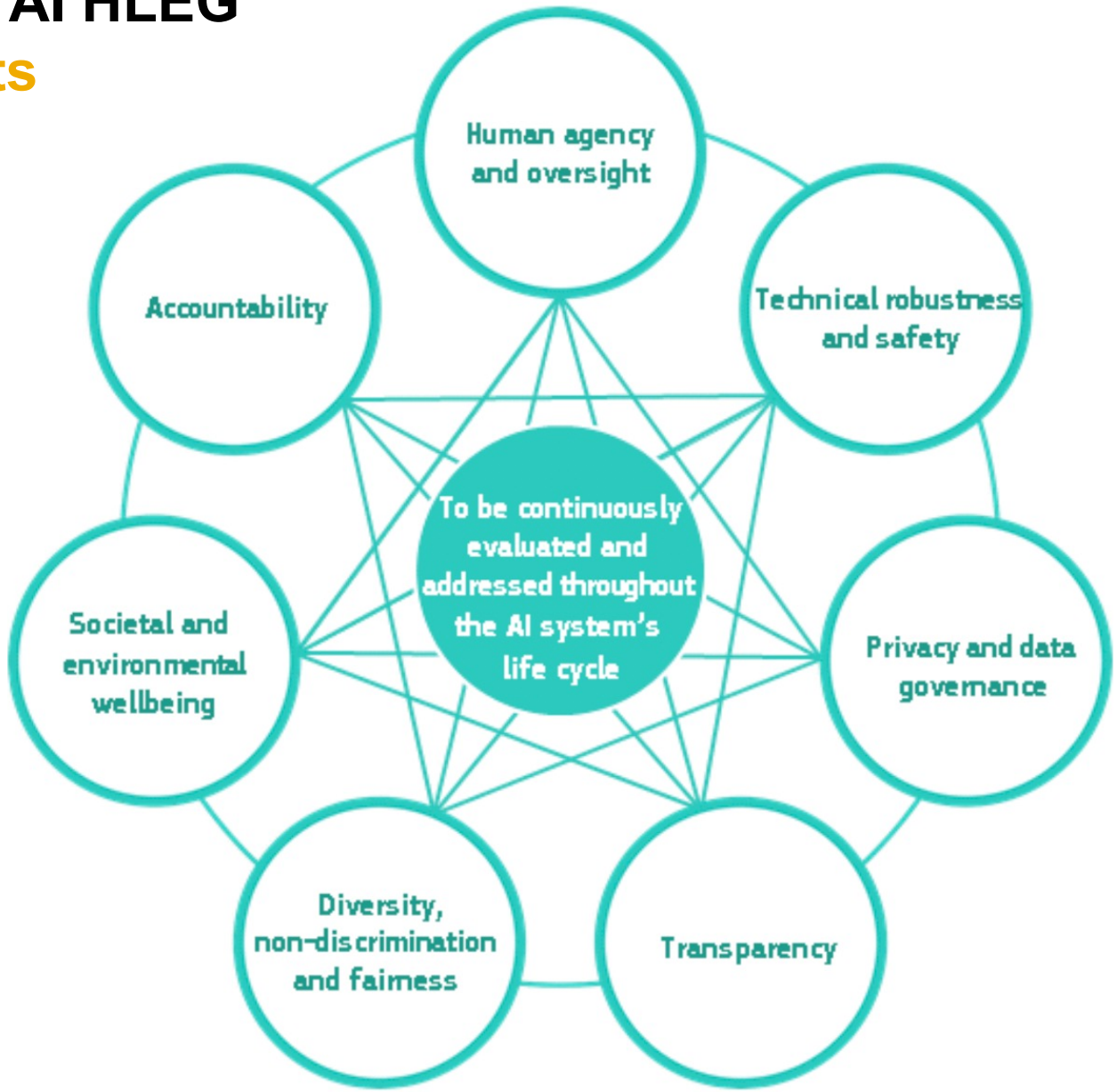
AI systems should ensure equal and just distribution of both benefits and costs, and ensure that individuals, and groups are free from unfair bias, discrimination and stigmatisation.

EXPLICABILITY

Processes need to be transparent, the capabilities and the purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected.

Trustworthy AI - AI HLEG

Key requirements



Spotlight on the EU AI Act

Scope and objectives

The **scope** of the Act is very wide, covering systems developed with any of the approaches in Annex I (machine learning, logic and knowledge-based approaches, and statistical or Bayesian approaches) that can generate outputs such as content, predictions, recommendations, or decisions influencing 'environments they interact with.

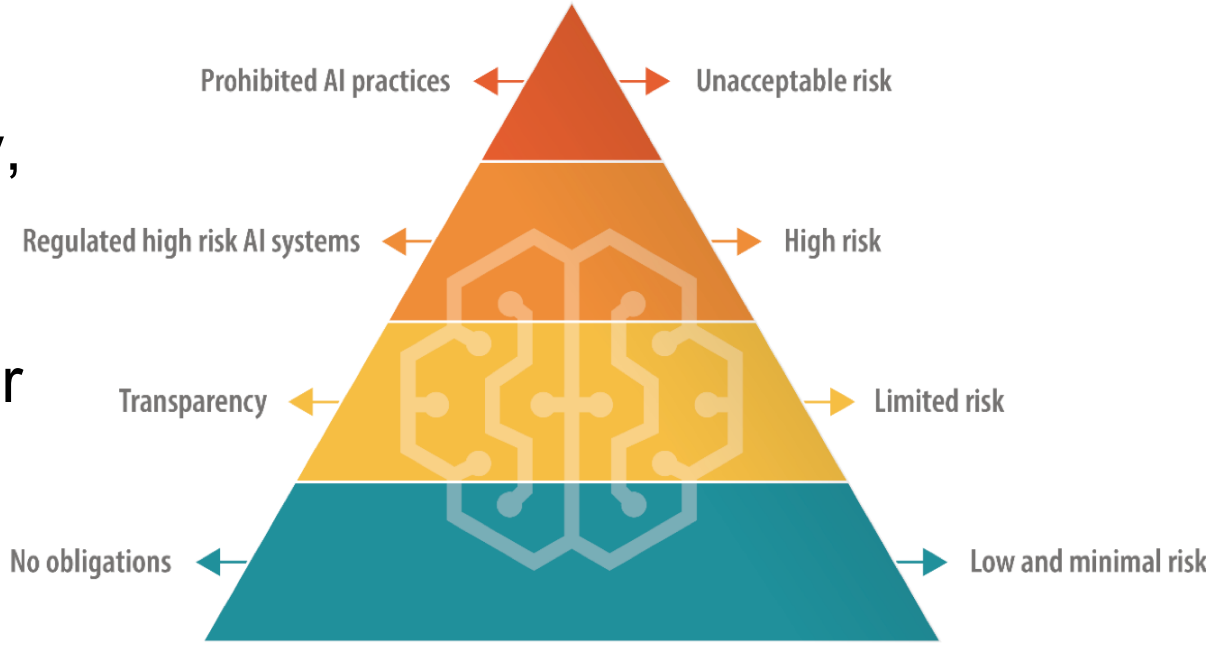
Objectives

- Ensure that AI systems placed on the Union market and used are safe and respect existing law on fundamental rights and Union values;
- Ensure legal certainty to facilitate investment and innovation in AI;
- Enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems;
- Facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation.

Spotlight on the EU AI Act

Risk-based approach

- **Prohibited use cases**: “Real-time” remote biometric identification systems, biometric categorisation systems using sensitive characteristics predictive policing systems, emotion recognition systems; untargeted scraping of facial images
- **High risk AI:** AI systems that pose significant harm to people’s health, safety, fundamental rights or the environment, AI systems used to influence voters and the outcome of elections and in recommender systems used by social media platforms



Spotlight on the EU AI Act

Recent updates

Obligations for general purpose AI

- Providers of foundation models would have to assess and mitigate possible risks and register their models in the EU database before their release on the EU market.
- Generative AI systems would have to comply with transparency requirements and ensure safeguards against generating illegal content.
- Detailed summaries of the copyrighted data used for their training would also have to be made publicly available.

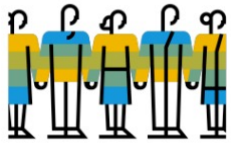
Supporting innovation and protecting citizens' rights

- The new law added exemptions for research activities and AI components provided under open-source licenses.
- The new law promotes so-called regulatory sandboxes, or real-life environments, established by public authorities to test AI before it is deployed.
- The new law boosts citizens’ right to file complaints about AI systems and receive explanations of decisions based on high-risk AI systems that significantly impact their fundamental rights.

Regulating AI

A real-world example with SAP

SAP's AI Ethics guiding principles



1. We are driven by our values.

2. We design for people.

3. We enable business beyond bias.

4. We strive for transparency and integrity.

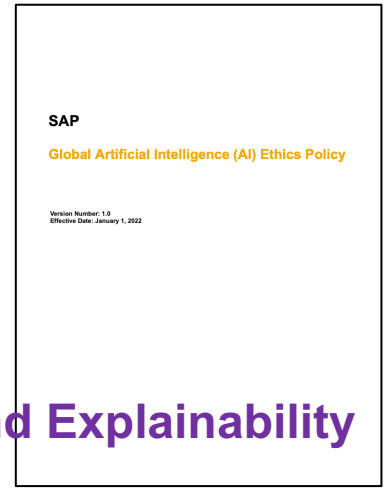
5. We uphold quality and safety standards.

6. We place data protection and privacy at our core.

7. We engage with the wider societal challenges of AI.

SAP's Guiding Principles for Artificial Intelligence: https://www.sap.com/france/documents/2018/09/940c6047-1c7d-0010-87a3-c30de2ffd8ff.html

SAP Global AI ethics policy
Framework



01. Human agency and oversight

Safeguarding human autonomy, particularly for automated decision-making.

- Enable humans to overrule decisions of AI system,
- Choose appropriate governance,
- Avoid unintended behavior.

02. Addressing Bias and Discrimination

Patterns of marginalization, inequality, and discrimination must not be encoded into AI.

- Build fair and unbiased AI systems,
- Use inclusive data for training,
- Realize measures to detect bias.

03. Transparency and Explainability

Prioritizing both the transparency of the development process and an AI‘s decisions.

- Document data sets and the development processes,
- Document the AI system's capabilities and limitations,
- Provide transparency about AI-generated output.

SAP Global AI ethics policy: https://www.sap.com/documents/2022/01/a8431b91-117e-0010-bca6-c68f7e60039b.html

SAP AI ethics handbook

Stakeholders of AI Factory Collaboration Process

AI Use Case Owner

Responsible for the coordination of the delivery of an AI enabled product or feature end to end across all phases of the AI factory process.

Factory Process Activity:

AI Data Scientist

Develops, implements or applies AI methods to derive solutions to business problems that can be translated into AI functions.

Factory Process Activity:

AI Engineer

Responsible for the design, implementation and maintenance of AI functions of a product.
Works according to specifications and project plans.

Factory Process Activity:

Product Developer

Develops the business logic of a product or at least parts of it. This task includes the integration of the provided AI functions into the product context.

Factory Process Activity:

Designer

Provides the Human-Centered understanding of the product by developing a user need strategy including stakeholder, end users, and product roadmap.
Conducts: user research, flows, usability evaluation, including voice response and conversational user interaction.

Factory Process Activity:

AI OPS Engineer

Operates the AI enabled products or individual AI functions and manages life-cycle aspects like versio updates.

Factory Process Activity:

User Assistance Developer

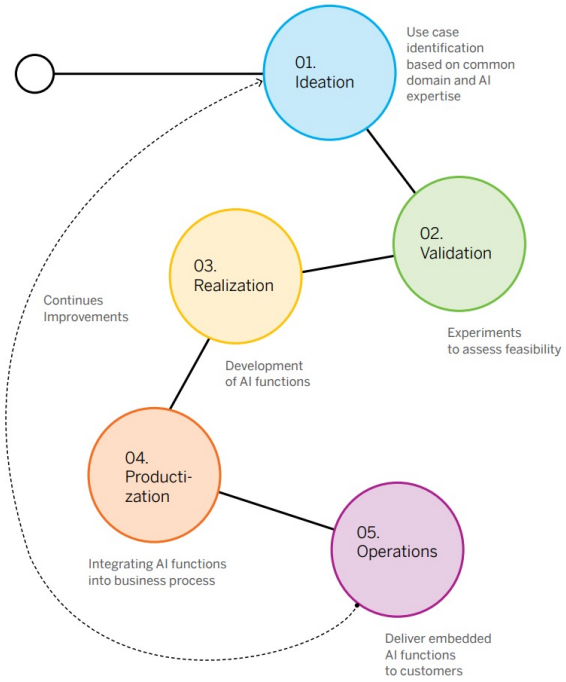
Responsible for designing, developing, reviewing, and maintaining content for product documentation, user interface messages, and conversational applications.
Collaborates with internal stakeholders to design customer-facing content and implements embedded h delivery mechanisms that improve how technical infor tion is delivered to customers.

Factory Process Activity:

Customer

Represents the user of the AI enabled product.
As user he must be active in the ideation, validation, ar operations phases.

Factory Process Activity:



SAP AI Ethics Handbook: https://www.sap.com/documents/2023/03/7211ee96-647e-0010-bca6-c68f7e60039b.html

Ethical AI governance at SAP

AI Ethics Steering committee

Senior leaders from design, data protection, corporate strategy, HR, sustainability, and legal with our AI researchers and operational specialists.

- Develops and enforces our guiding principles.
- Assesses high_risk use cases and provides guidance to use case owners.

Provides updates on activities

Provides updates on activities

Provides guidance

AI Ethics Advisory Panel

SAP external experts with a public mandate from academia, politics, and industry.

- Provides input on the guiding principles
- Advises on the operationalization of the guiding principles

Trutworthy AI Workstream

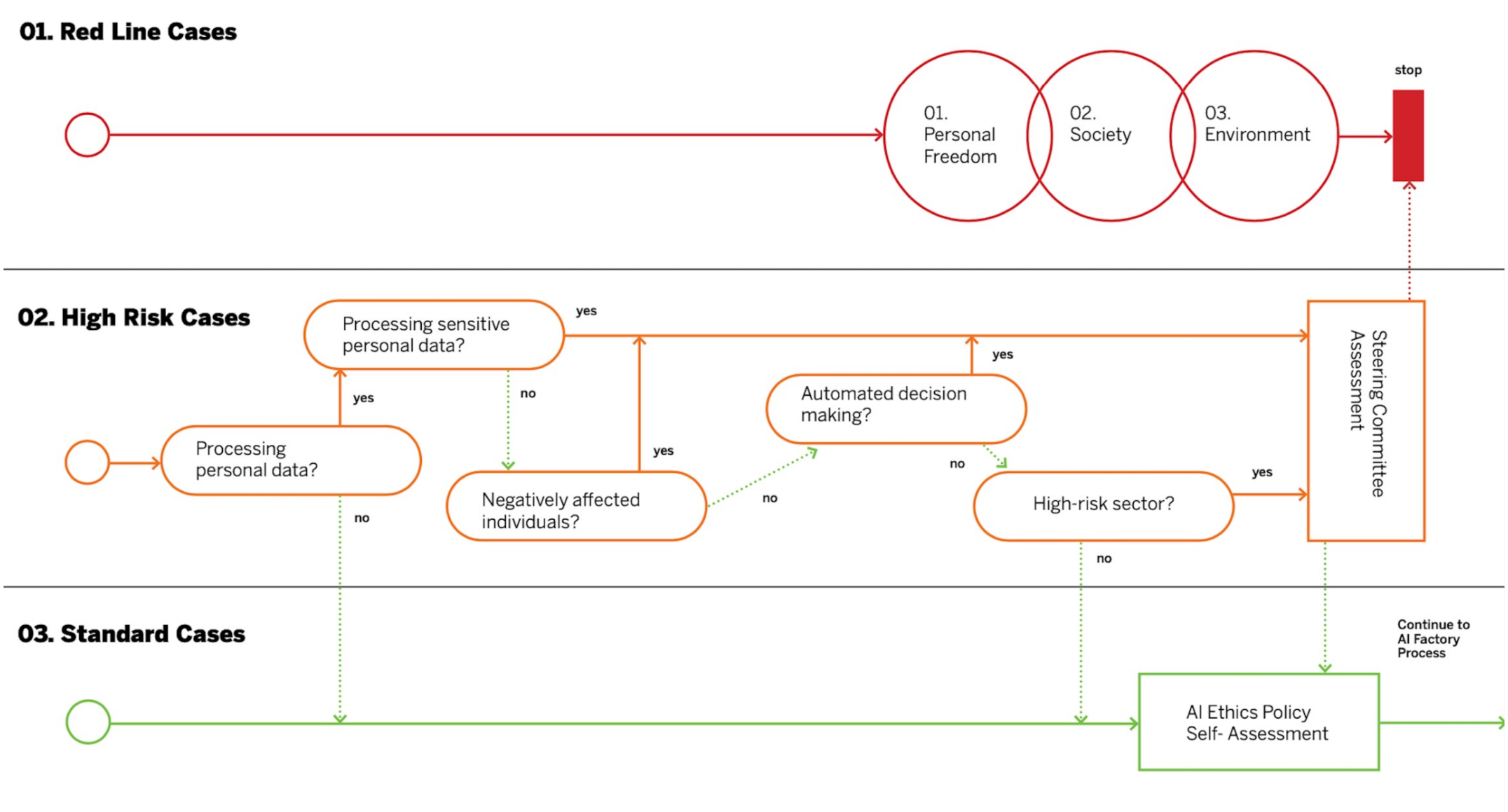
Group of interested SAP employees who want to engage with Trustworthy AI and build expertise

- Establishes the means to implement the necessary processes to ensure compliance in SAP's AI development.

SAP AI Ethics: https://www.sap.com/france/products/artificial-intelligence/ai-ethics.html

Risk classification and assessment process

Overview



On the technical challenges:

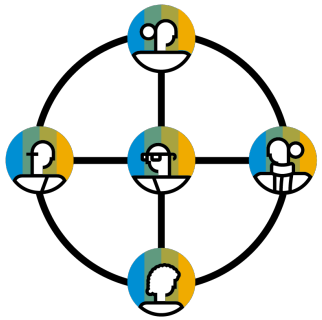
AI Fairness, explainability, privacy and security

AI Fairness

A fair AI model ensures **equitable treatment of individuals or groups** without bias, discrimination, or unfair advantage, aiming to mitigate and prevent the perpetuation or amplification of existing societal disparities.

Sources of bias

- Lack of data diversity and representativeness,
- Data collection process,
- Historical/social bias.

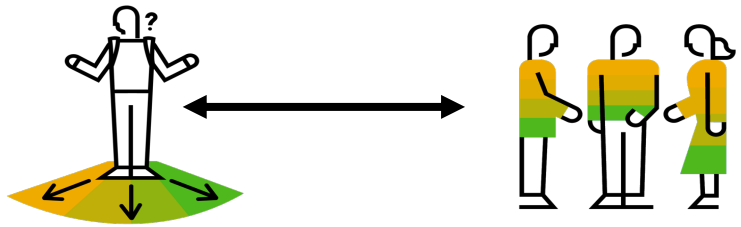


Types of bias

- Data bias
- Algorithmic bias

AI Fairness

Fairness through unawareness

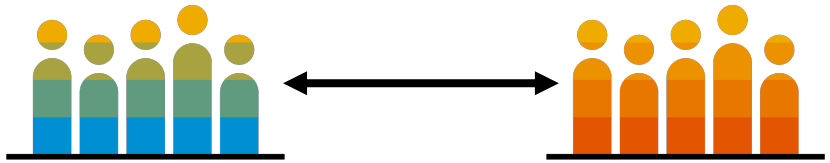


- Making decisions by deliberately discarding certain sensitive characteristics of individuals,
- Problem of proxy features, which are features that are indirectly related to a sensitive attribute and can be used to infer it.

Dwork, Cynthia, et al. "Fairness through awareness." *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012.

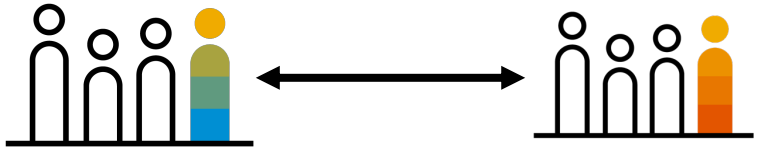
AI Fairness

Group fairness



- Different demographic groups must be treated equally
- Regardless of demographic group membership,
- Examines the differences between demographic groups and compares them with their predictions,
- Incompatibility in group fairness notions.

Individual fairness



- Similar people must be treated in the same way,
- Regardless of demographic group,
- Examines the differences between individuals and compares them with their predictions.
- Complexity and subjectivity in defining fairness at the individual level,
- Establishing a reliable mapping between individual similarity and prediction similarity is challenging.

AI Fairness – group fairness

Demographic parity

Equalized opportunity

- A predictor that satisfies the demographic parity constraint should yield that each subgroup receives a positive outcome at equal rates.
- Demographic parity can be expressed as follows, where $\hat{Y}$ is the predicted label and A is the protected attribute indicating the subgroup:

$$P(\hat{Y} = 1 | A = a) = P(\hat{Y} = 1 | A = b), \forall\, a, b \in A$$

- A predictor that satisfies the equalised opportunity constraint should yield that the true positive rate of each subgroup should be equal.
- Equalised opportunity can be expressed as follows, where A is the sensitive attribute, Y' is the predicted label and Y is the ground truth label:

$$P(\hat{Y} = 1 | Y = y1\ A = a) = P(\hat{Y} = 1 | Y = 1, A = b), \quad \forall a, b \in A$$

AI Fairness – group fairness

Demographic parity

Equalized opportunity

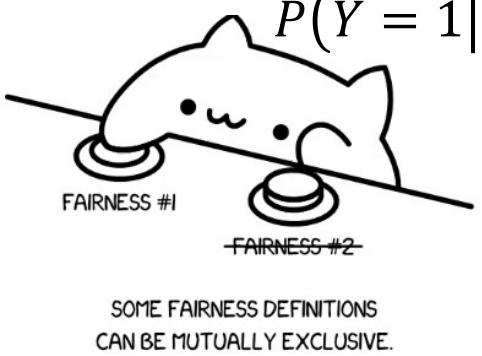
- A predictor that satisfies the demographic parity constraint should yield that each subgroup receives a positive outcome at equal rates.
- Demographic parity can be expressed as follows, where $\hat{Y}$ is the predicted label and A is the protected attribute indicating the subgroup:

$$P(\hat{Y} = 1 | A = a) = P(\hat{Y} = 1 | A = b), \forall\, a, b \in A$$

- A predictor that satisfies the equalised opportunity constraint should yield that the true positive rate of each subgroup should be equal.
- Equalised opportunity can be expressed as follows, where A is the sensitive attribute, Y' is the predicted label and Y is the ground truth label:

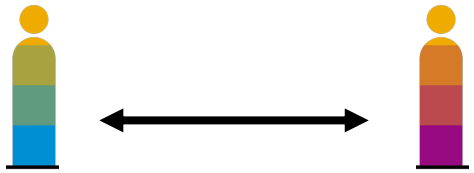
$$P(\hat{Y} = 1 | Y = y1\ A = a) = P(\hat{Y} = 1 | Y = 1, A = b),$$
$$\forall a, b \in A$$

Impossibility theorem



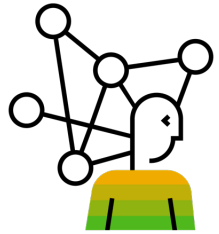
AI Fairness – individual fairness

Counterfactual fairness



- Evaluating fairness by considering counterfactual scenarios.
- Assessing whether individuals would receive fair treatment if their characteristics or attributes were different while holding other relevant factors constant.
- Access to reliable and comprehensive counterfactual data is challenging

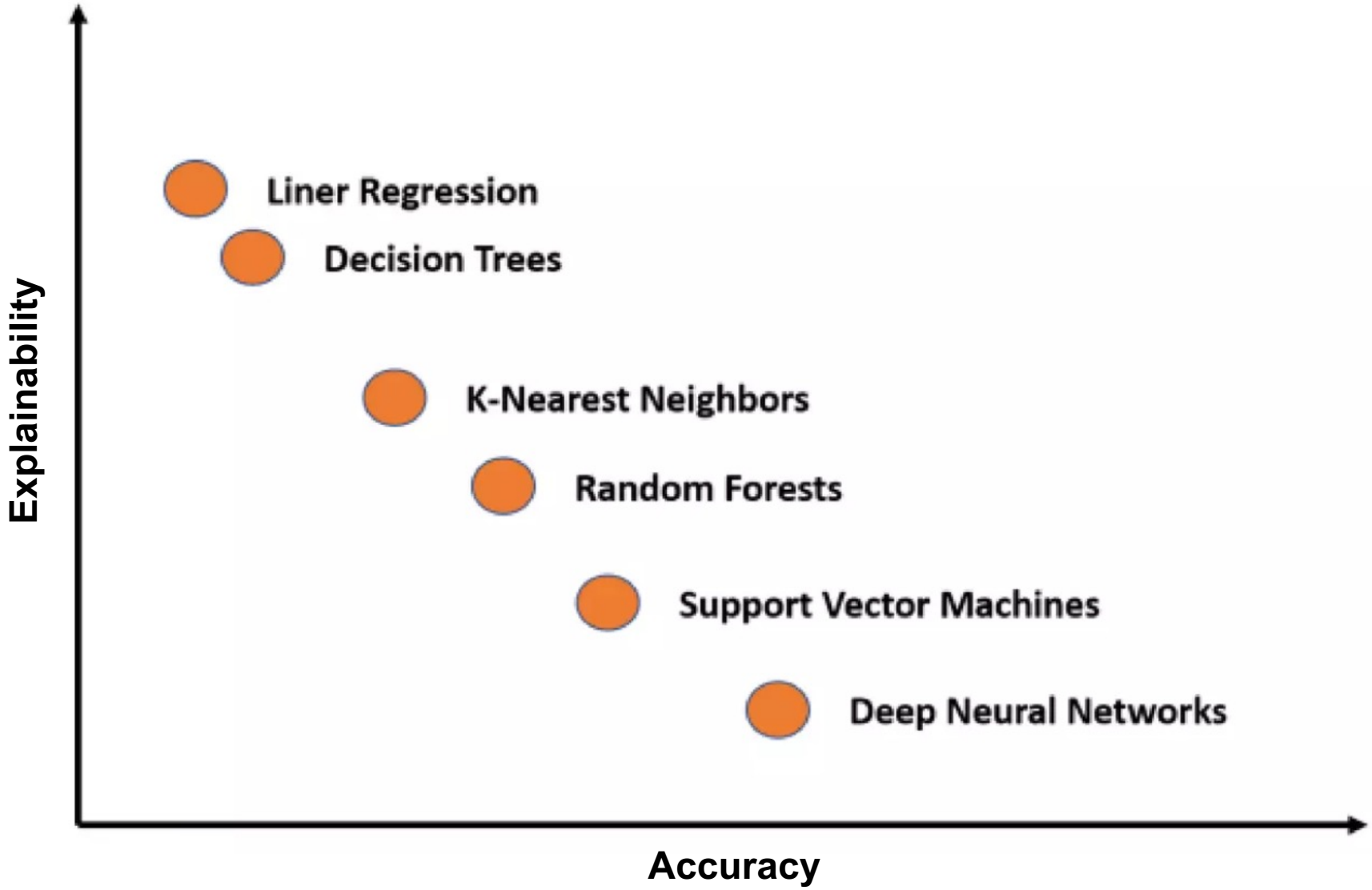
Causal fairness



- Evaluates fairness by considering causal relationships between attributes and decision outcomes.
- Focuses on understanding and addressing the underlying causes of disparities in individual outcomes.
- Identifying relevant causal factors, especially in real-world scenarios with confounding variables and indirect effects.

[1] Kusner, Matt J., et al. "Counterfactual fairness." *Advances in neural information processing systems* 30 (2017).
[2] Plecko, Drago, and Elias Bareinboim. "Causal fairness analysis." *arXiv preprint arXiv:2207.11385* (2022).

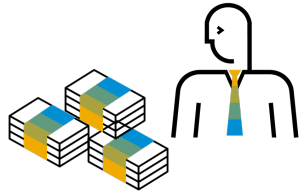
AI explainability



AI explainability

Why am I receiving this decision?

How can I get a better outcome?

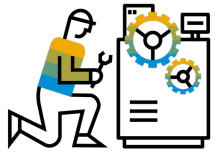


Business Owner

Can I trust our AI decision?

Costumer support

How do I respond to this customer's complaint?

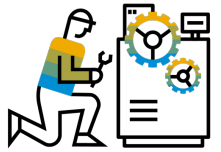


IT & operations

How can I control and debug this model?

Data scientist

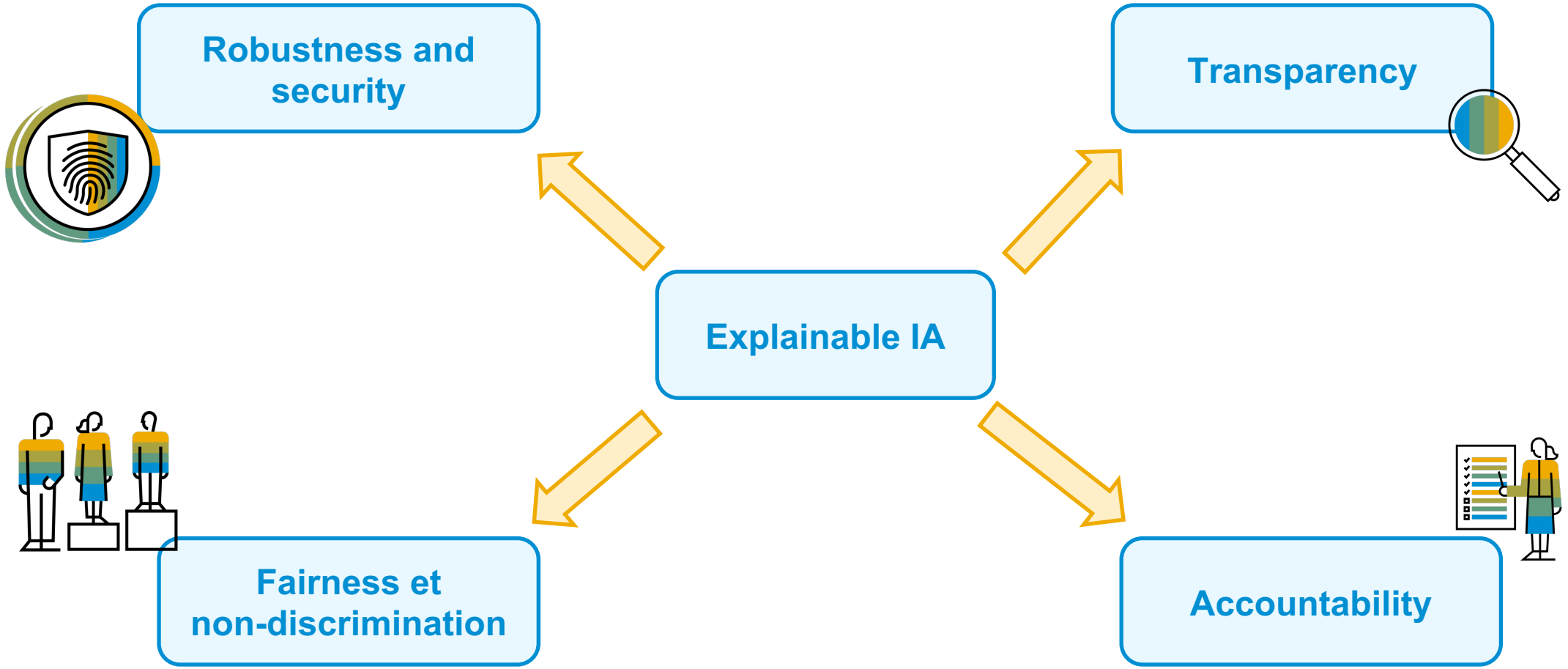
Is this the best model we can build?



Internal audits, regulators

Are these AI decisions harmless?

AI explainability



AI explainability

Global explanations

Explains overall AI model behavior.

Local explanations

Explains the AI model's prediction for each given case.

Model-agnostic explainability

Provides explanations of any AI model.

Model-specific explainability

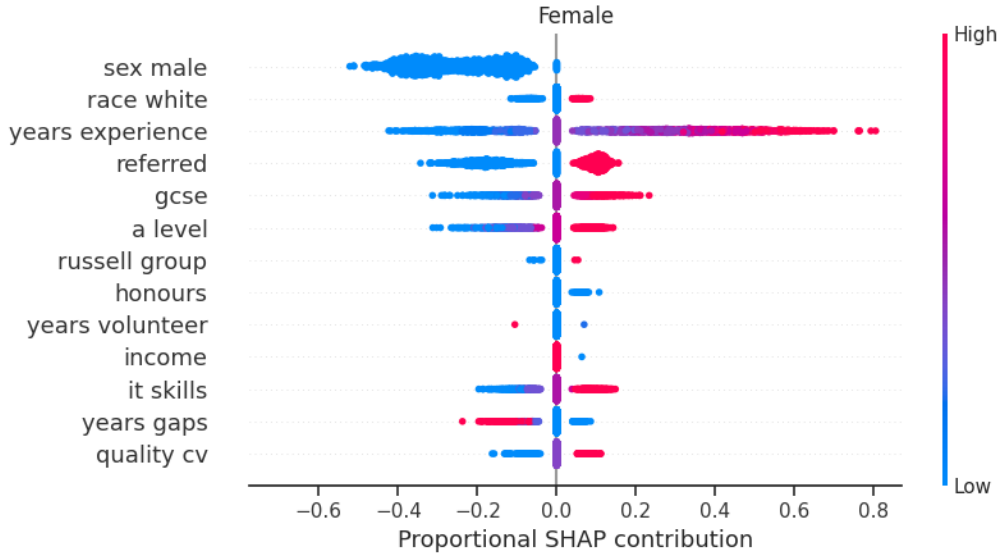
Explanations are based on the learning process of the given model.

AI explainability

SHAP

SHapley Additive exPlanations

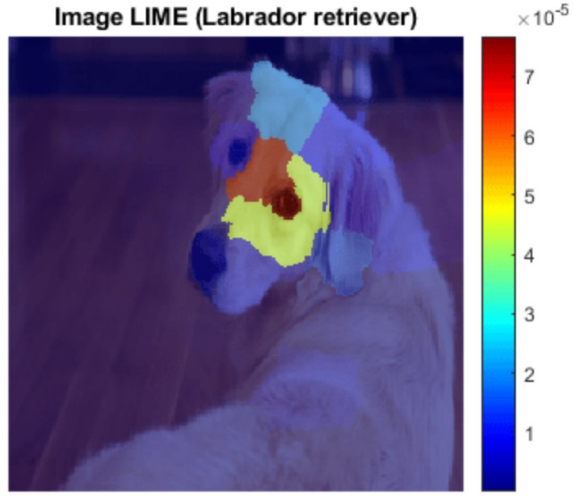
- SHAP works by assigning importance values to each attribute of an input instance by combining the marginal contributions of each attribute, based on game theory, to explain the predictions of a machine learning model.



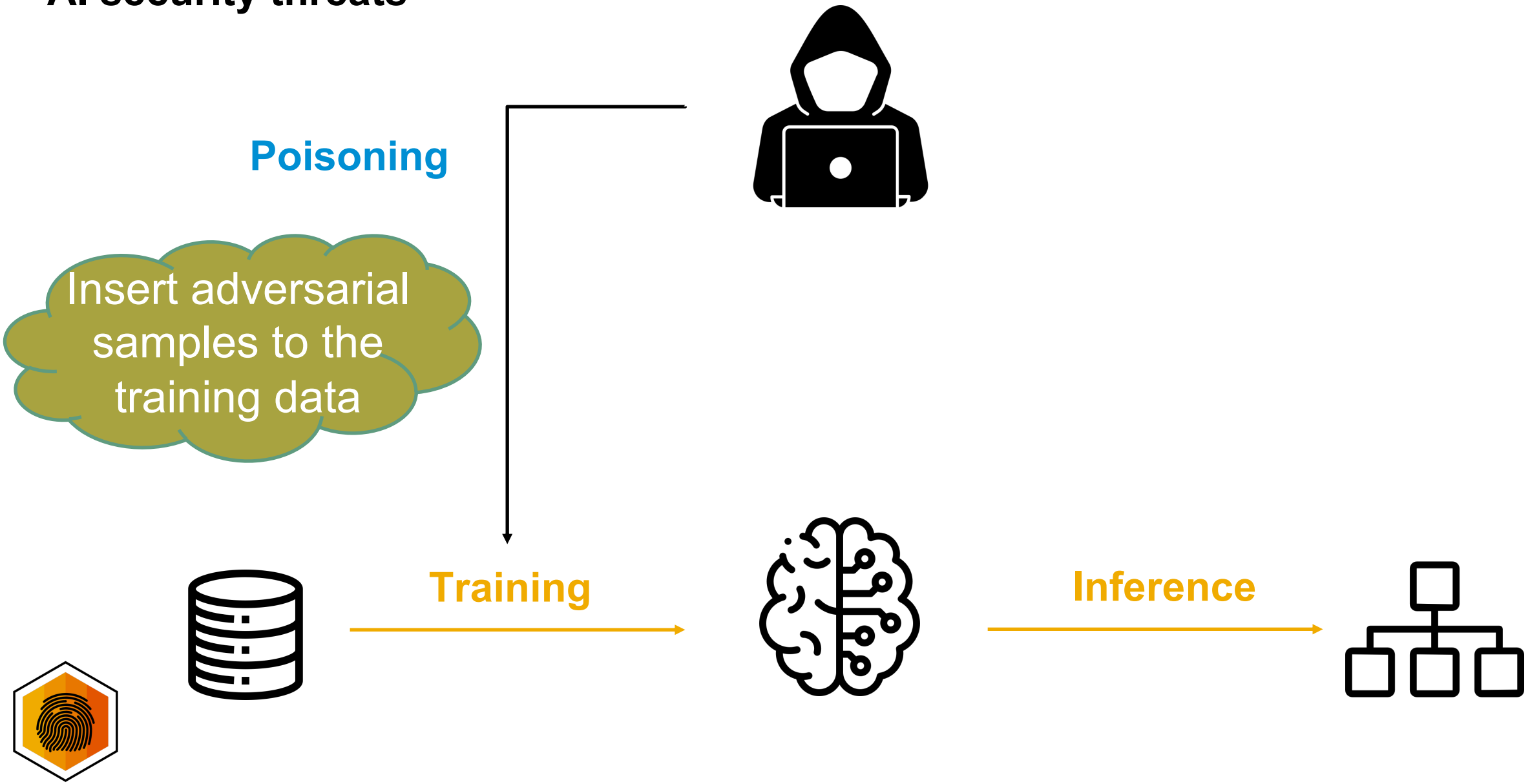
LIME

Local Interpretable Model-Agnostic Explanations

- Works by locally approximating the predictions of a complex machine learning model by building an interpretable linear model on a restricted neighborhood around a given input instance.
- LIME uses attribute perturbation to evaluate the impact of each attribute on model prediction.



AI security threats



AI security threats

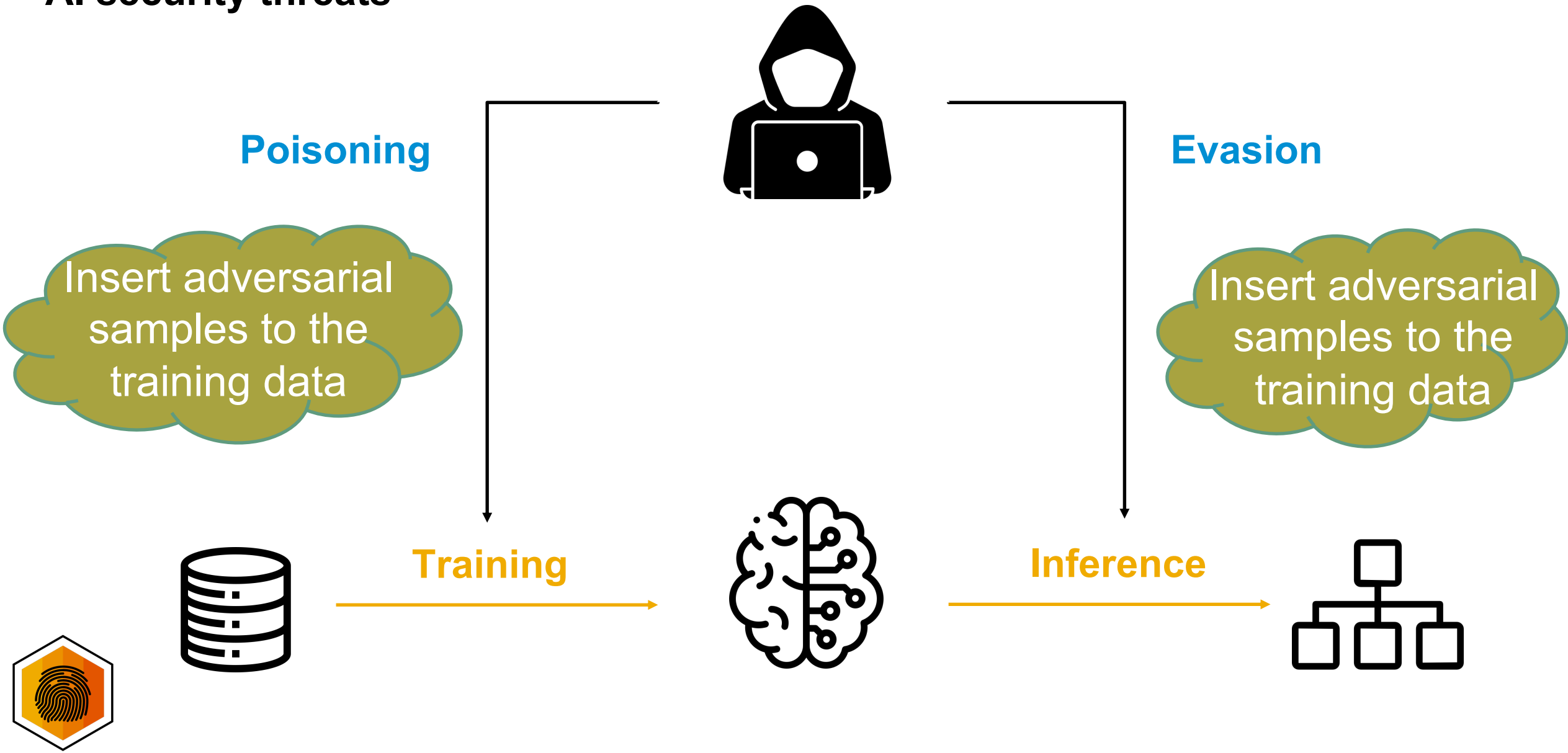
Data poisoning attack

- **What is it?** AI model poisoning is an attack where an adversary manipulates the training data to insert malicious samples.
- **Impact:** This can lead to AI models making incorrect or harmful predictions.
- **Examples:** Email spam filtering model

AI security threats
Backdoor attacks

- **What is it?** Backdoor attacks involve inserting hidden triggers or patterns into the training data, which can be activated to manipulate the model's behaviour.
- **Impact:** Attackers can compromise the AI system's integrity and make it behave maliciously.
- **Examples:** Inserting a given pattern that will trigger the loan approval automatically in a bank’s Credit Risk Assessment Models.

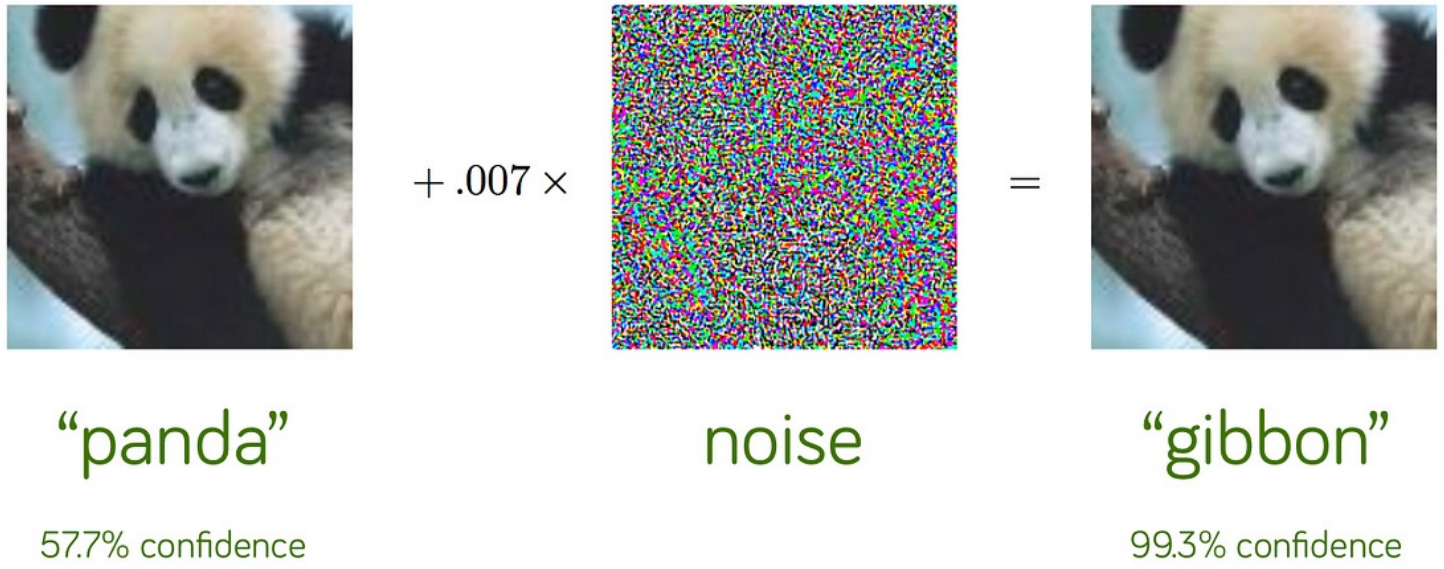
AI security threats



AI security threats

Evasion attacks - Adversarial attacks

- **What is it?** Evasion attacks involve modifying input data to fool an AI model.
- **Impact**: Attackers can make AI systems misclassify or ignore specific inputs.

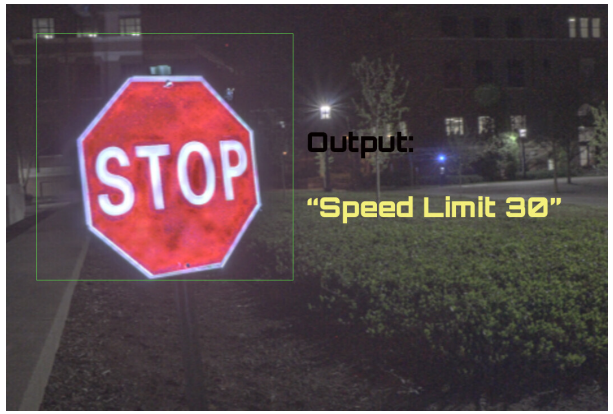


AI security threats

Evasion attacks - Adversarial attacks

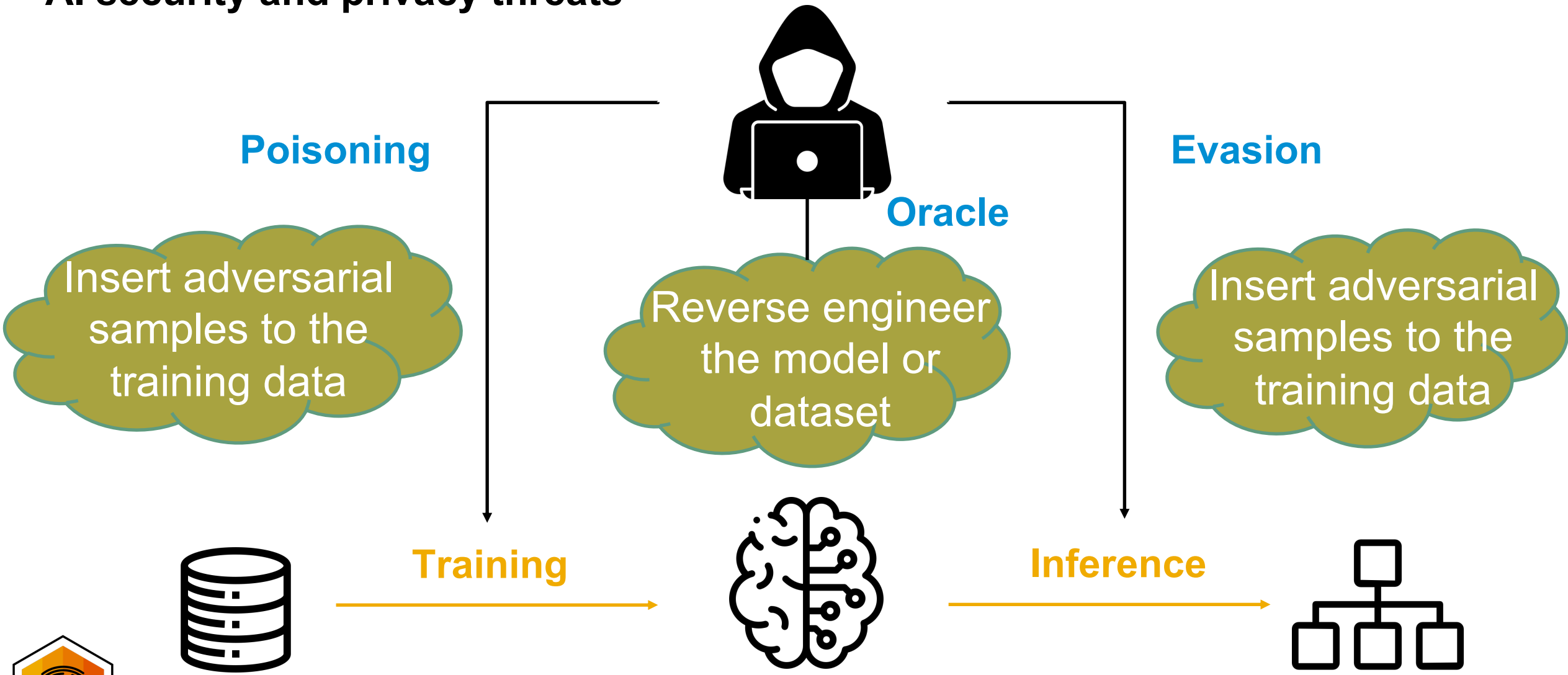
- **What is it?** Evasion attacks involve modifying input data to fool an AI model.
- **Impact**: Attackers can make AI systems misclassify or ignore specific inputs.
- **Examples:** Maliciously crafted images can trick AI-based facial recognition systems, autonomous vehicles, etc.

Masks that fool facial recongition technology



Researchers fooled AI into ignoring stop signs using a cheap projector

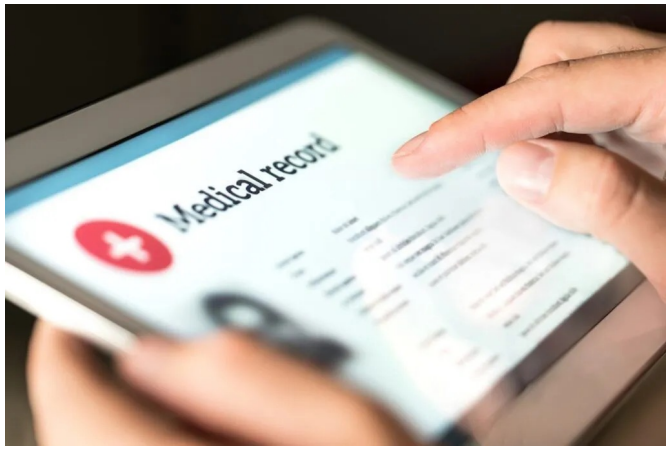
AI security and privacy threats



AI privacy threats

Membership inference attacks

- **What is it?** Membership inference attacks aim to determine whether a data point from a specific individual has been used during the training of the AI model.
- **Impact**: Attackers can infer sensitive information by identifying training data.
- **Examples:** Identifying whether an individual's health record was part of a medical AI training dataset.



AI privacy threats
Model inversion attacks

- **What is it?** Model inversion attacks aim to reconstruct sensitive information about individuals in the training data by inverting the original model using input-output pairs.
- **Impact**: Attackers can infer personal and sensitive data, potentially violating privacy regulations.
- **Examples:** Reveal sensitive information about someone's drug consumption past by inverting a model intended for anti-drug awareness.

AI privacy threats

IP theft - Model extraction attacks

- **What is it?** Model extraction attacks involve reverse-engineering an AI model based on its outputs.
- **Impact:** Attackers can clone proprietary AI models, potentially stealing intellectual property.
- **Examples:** Extracting Spotify's music recommendation model to create a competing music platform offering competitive pricing.

MITRE | ATLAS™ Matrices Navigator Tactics Techniques Mitigations Case Studies Resources

Draft of ATLAS mitigations are now available - feedback welcome!

MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems), is a knowledge base of adversary tactics, techniques, and case studies for machine learning (ML) systems based on real-world observations, demonstrations from ML red teams and security groups, and the state of the possible from academic research. ATLAS is modeled after the MITRE ATT&CK® framework and its tactics and techniques are complementary to those in ATT&CK.

ATLAS enables researchers to navigate the landscape of threats to machine learning systems . ML is increasingly used across a variety of industries. There are a growing number of vulnerabilities in ML, and its use increases the attack surface of existing systems. We developed ATLAS to raise awareness of these threats and present them in a way familiar to security researchers.

ATLAS™

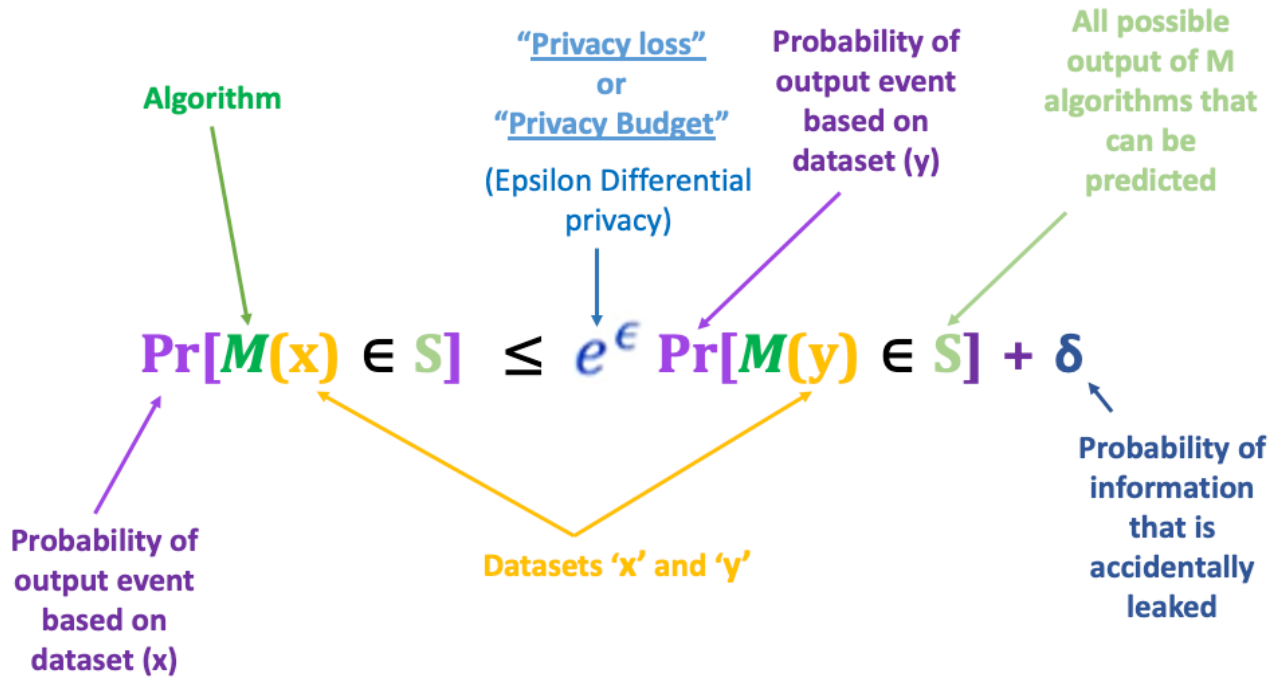
The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below.
& indicates an adaptation from ATT&CK. Click on links to learn more about each item, or view ATLAS tactics and techniques using the links at the top navigation bar.

Reconnaissance &	Resource Development &	Initial Access &	ML Model Access	Execution &	Persistence &	Defense Evasion &	Discovery &	Collection &	ML Attack Staging	Exfiltration &	Impact &
5 techniques	7 techniques	4 techniques	4 techniques	2 techniques	2 techniques	1 technique	3 techniques	3 techniques	4 techniques	2 techniques	7 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution &	Poison Training Data	Evade ML Model	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
	Develop Adversarial ML Attack Capabilities	Evade ML Model	Physical Environment Access				Discover ML Artifacts	Data from Local System &	Verify Attack		Spamming ML System with Chaff Data
Owned Websites	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model						Craft Adversarial Data		Erode ML Model

Tutorial:
Fair and Private Machine Learning

Differential privacy

- Anonymization technique,
- mathematical guarantee that any outcome of a data analysis cannot be tied back to any individual's data in a dataset,
- Utility-privacy tradeoff,
- Fairness-privacy tradeoff.



[1] Dwork, C. (2006). Differential Privacy. Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP '06), 1-12.
[2] Noe, F., Herskind, R., and Søgaard, A. (2022). Exploring the unfairness of DP-SGD across settings. *CoRR*, abs/2202.12058.
[3] Xu, D., Du, W., and Wu, X. (2020). Removing disparate impact of differentially private stochastic gradient descent on model accuracy. *CoRR*, abs/2003.03699.
[4] Ganev, G., Oprisanu, B., and De Cristofaro, E. (2022). Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6944–6959. PMLR.
[5] Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., and Miklau, G. (2020). Fair decision making using privacy-protected data. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 189–199. ACM.
[6] Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019*.

Tutorial

Fair and Private Machine Learning

IBM/**differential-privacy-library**

IBM

Tutorial link:

https://colab.research.google.com/drive/1qCL3qR5VHfnJtMGNITI7akvzqGojkUuU?usp=sharing

Thank you.

Contact information:

Khawla Mallat
Research Scientist – Trustworthy AI
khawla.mallat@sap.com