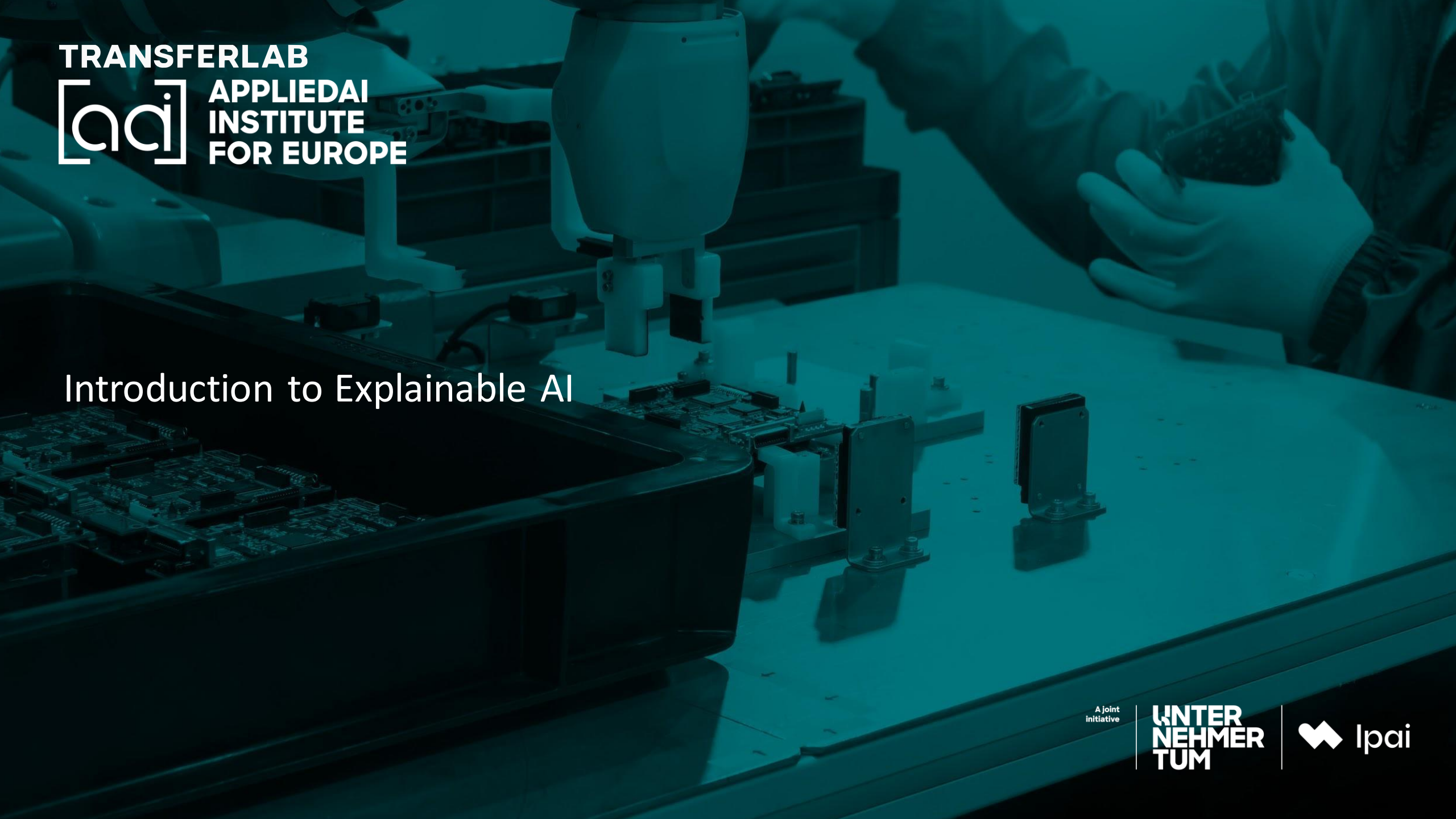


Introduction to Explainable AI

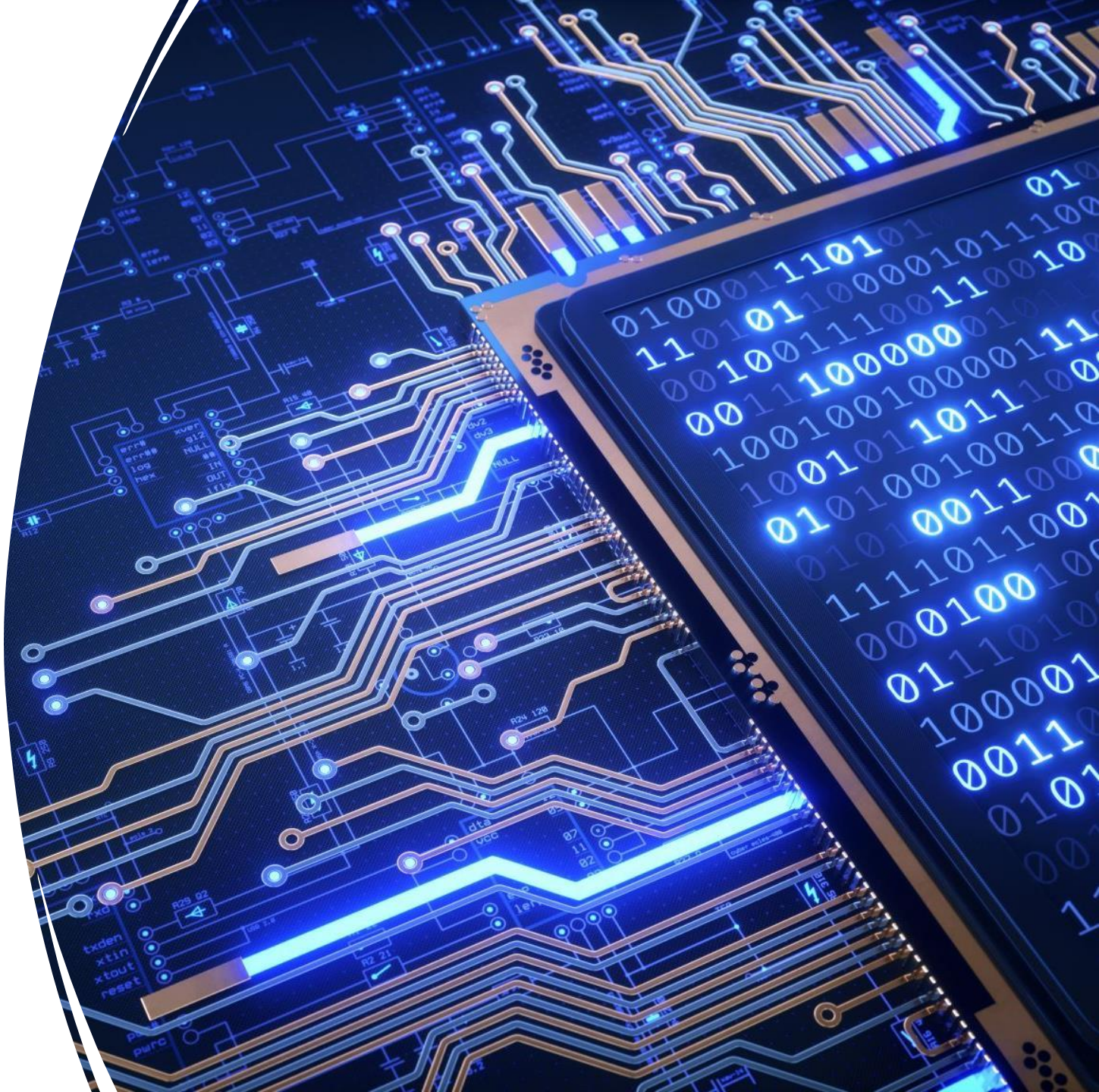
A joint initiative

UNTERNEHMER TUM

Ipai

Impact of Artificial Intelligence

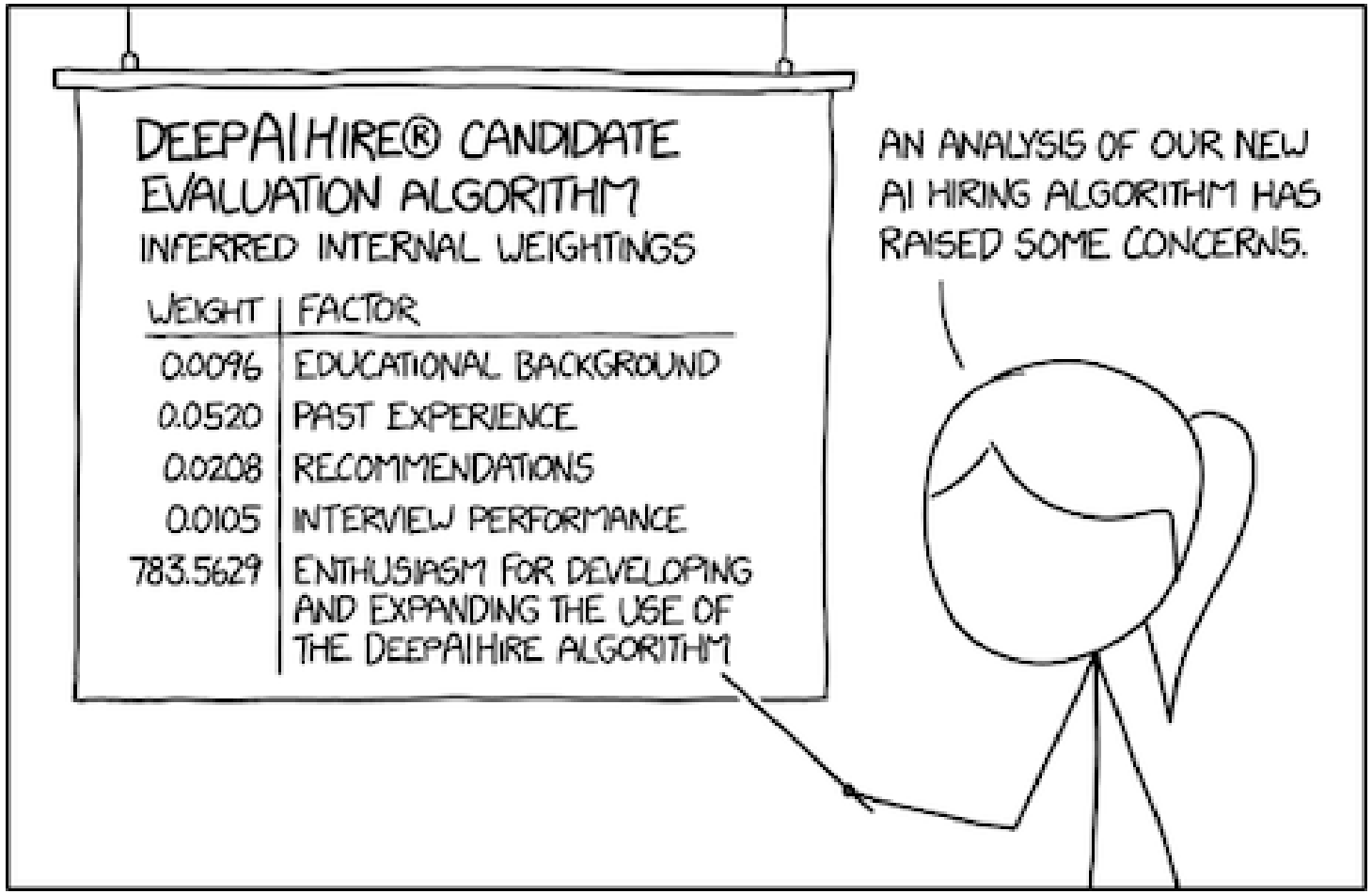
- The adoption of artificial intelligence into the professional and private live is happening with increasing pace.
- As permeation of the society progresses, the number of incidents naturally also rise.
- The AI incidents database keeps track of such incidents.



Bias

Machine learning systems are subject to numerous sources for bias, e.g.

- Induced by the architecture/algorithm
- Contained in the data, e.g., selection bias
- Contained in the labels, e.g., social bias
- ...



Discrimination Related Incidents

A bias in a machine learning model is particularly problematic if it supports discrimination.

Discrimination Incidents

Criminal Risk Assessment

An analysis of over 10000 criminal defendants in Broward County, Florida. Black defendants were often predicted to be at a higher risk of recidivism than they were. Black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent, link).

Safety

- As the stakes behind AI driven decisions get higher, the consequences of a malfunction become more severe.
- Autonomous systems such as assembly robots or self-driving cars can create huge forces.
- A malfunction can cause very severe accidents.
- A model that is overfitted and relies on too simple patterns might fail when confronted with rare events or distribution shift at inference time.
- Explainability methods can help to understand how a model will behave even in cases where no data is available.

Safety Related Incidents

Safety Related Incidents

Paris Taxi Accident

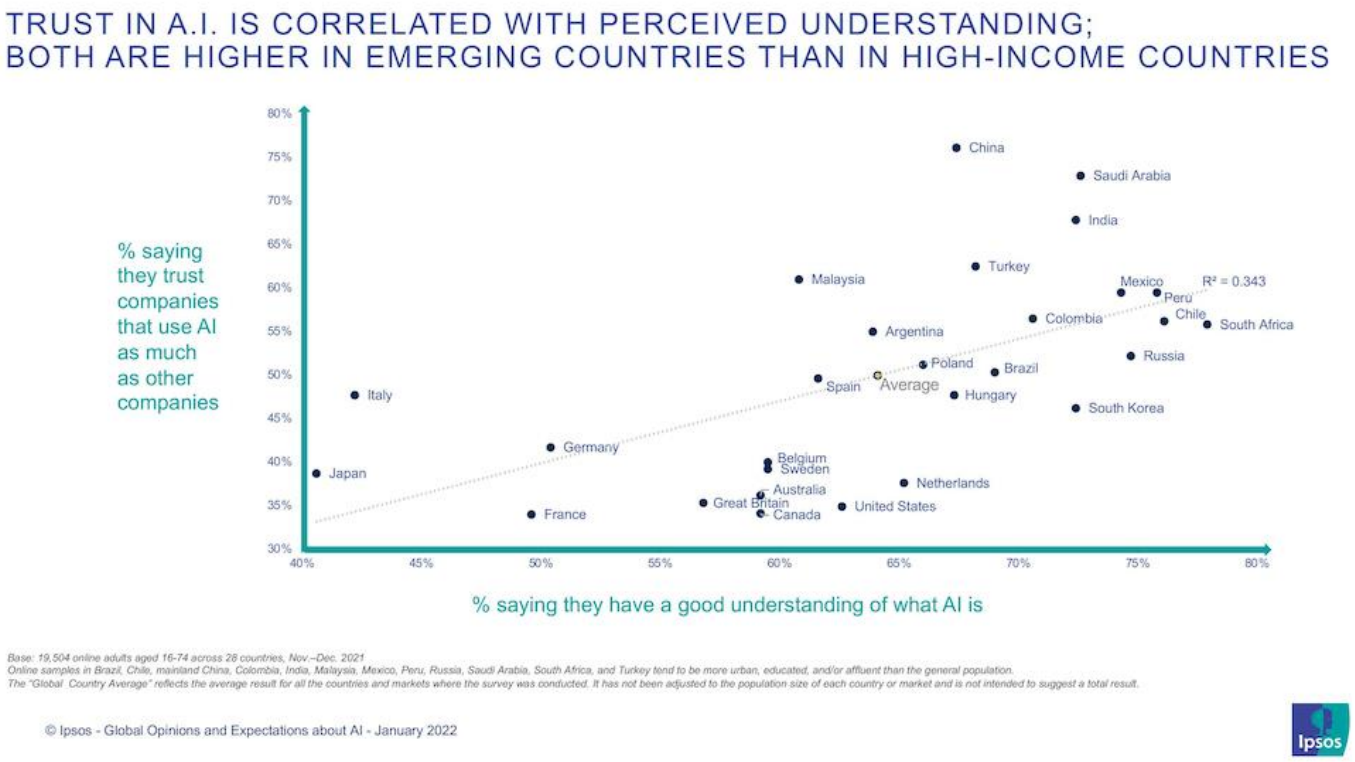
A Tesla Model 3 taxicab got involved in a severe accident in Paris. The first official information about it is that the cause was an SUA (sudden unintended acceleration) episode and braking issues. About 20 people got hurt, five of them with life-threatening injuries (2021, link).

The Blackbox Problem

- Many machine learning models derive their predictions through a process that is hard to interpret.
- Machine learning models are therefore Backboxes to the human understanding.
- Although the model might perform well, it does not allow to draw conclusions on why it made a certain prediction.
- This can cause serious trust issues, especially for high-stake decisions like credit-loans or medical diagnosis.

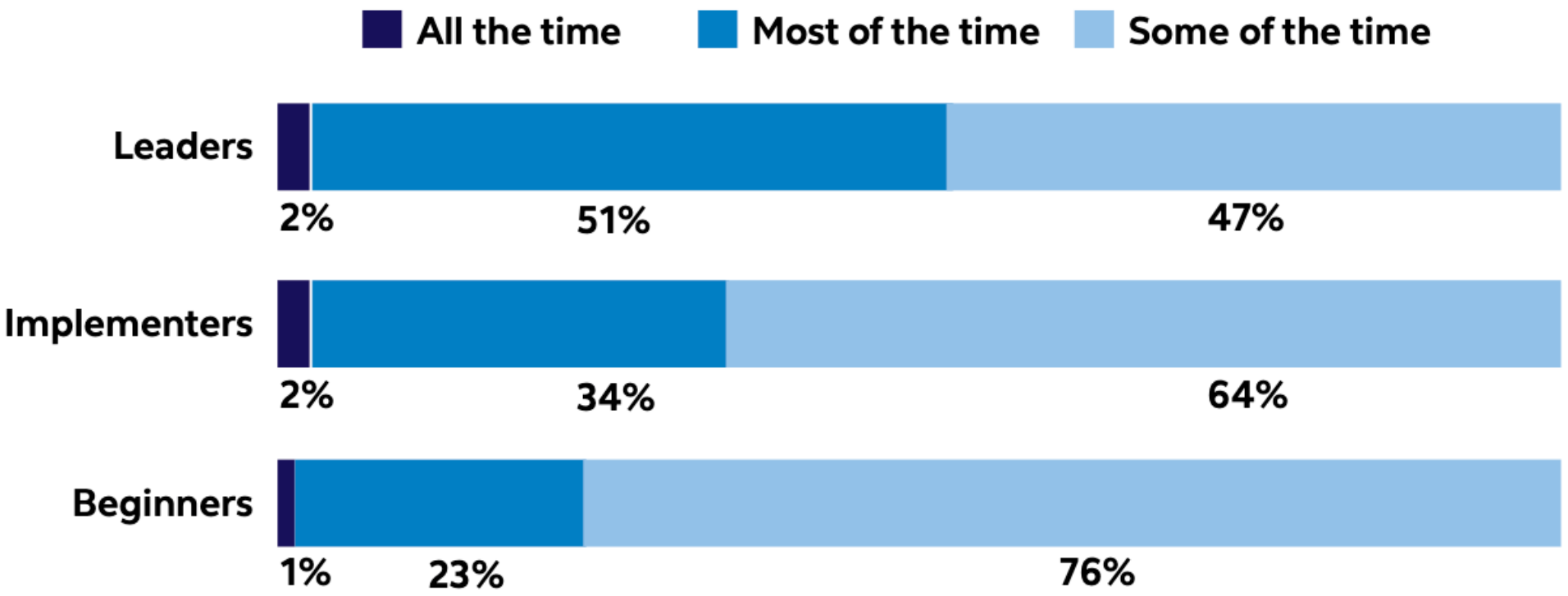
Trust in AI

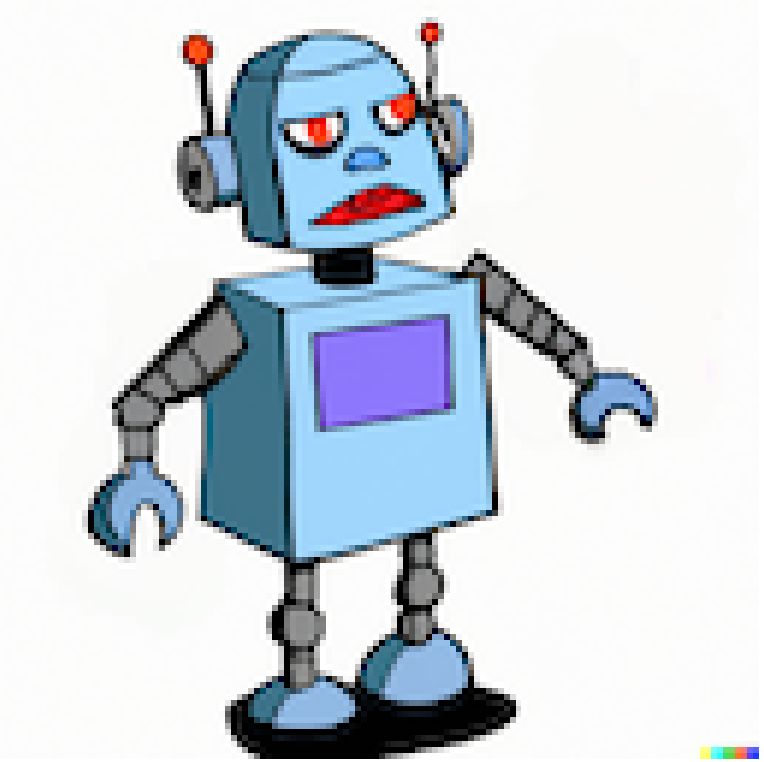
Online survey by the world economic forum (link)



Do Executives Trust their AI Systems?

Taken from a 2021 survey by Cognizant among 1000 senior executives from a broad range of US industries. (link)



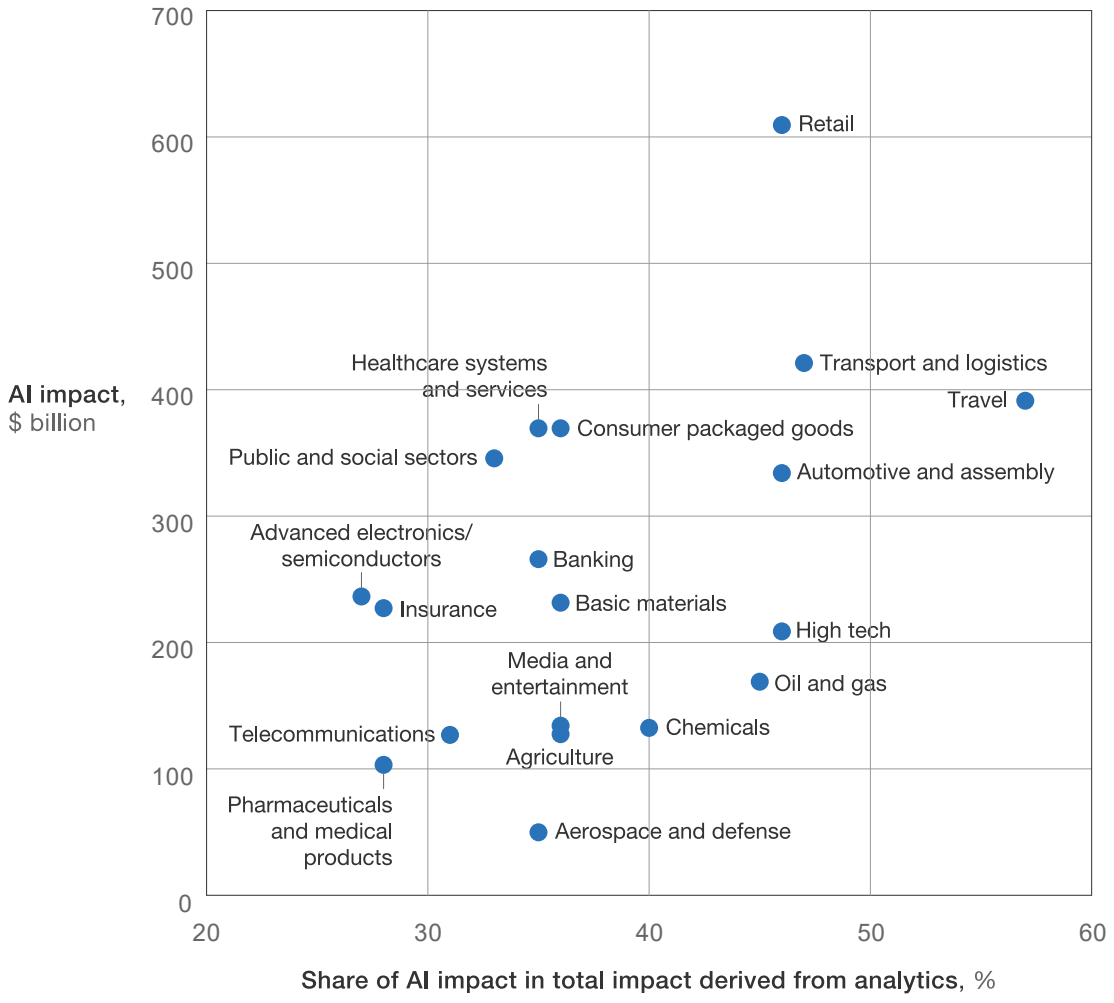


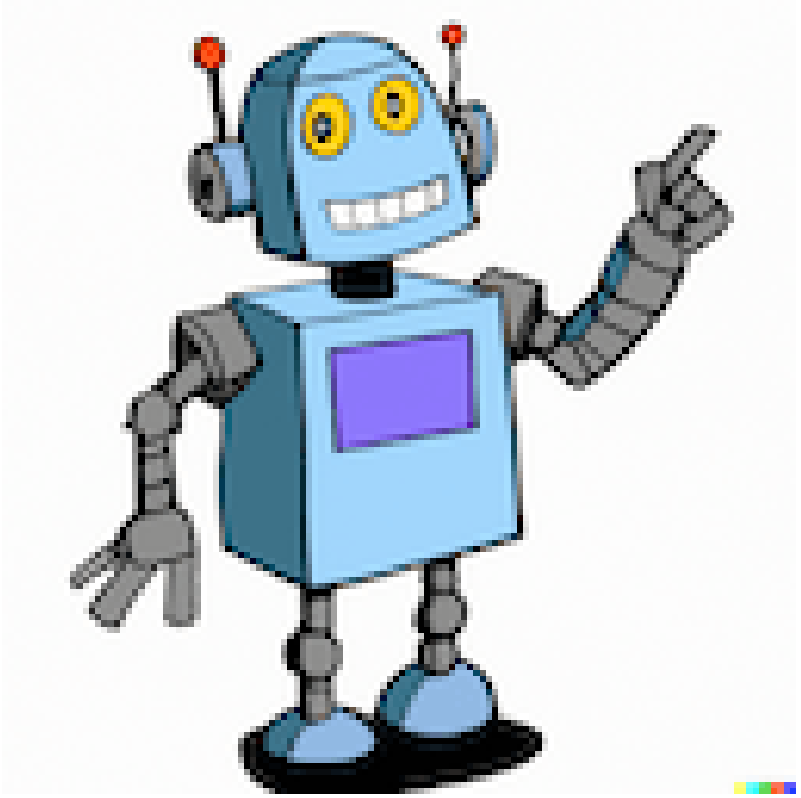
Distrust

- Distrust can be a major barrier for the adoption of AI.
- This can lead to high opportunity costs and systematic competitive disadvantages due to late adoption.

AI Potential

Taken from 2018 study by McKinsey based on several hundred use-cases. They identified together between $3.5 trillion and $5.8 trillion in value annually across nine business functions in 19 industries (link).





Obedience

- Some people are too naive towards decisions taken by autonomous systems.
- They are likely to accept predictions without much skepticism.

Navigation Systems

Yet Another Person Listens to GPS App and Drives Car Into Lake. Some tourists driving through Vermont said they ended up in a lake and blame the driving app Waze (2018).

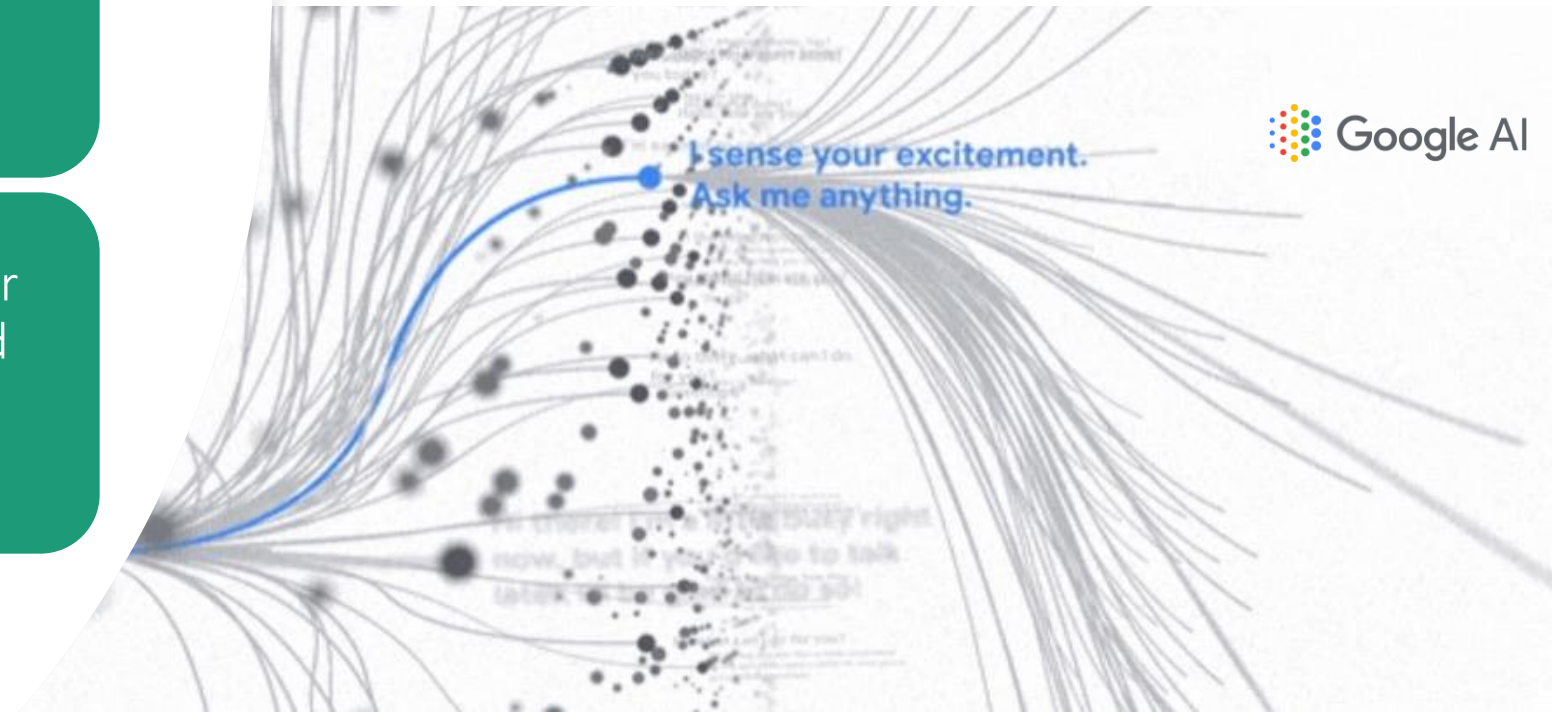
Conscious Machines?

[Eliza](): A simple rephrasing-based chat program (psychotherapy) from 1966. After typing a conversation, many test users where convinced that the machine understood their concerns ([demo]()).

[LaMDA](): A Google engineer was fired after publicly stating that the new GPT-3 based chat service LaMDA has developed a conciousness.

```
Eliza is a mock Rogerian psychotherapist.
e original program was described by Joseph Weizenbaum in 1966.
s implementation by Norbert Landsteiner 2005.

s something troubling you ?
n are all alike.
t is the connection, do you suppose ?
y're always bugging us about something or other.
you think of a specific example ?
, my boyfriend made me come here.
important to you that your boyfriend made you come here ?
ys I'm depressed much of the time.
sorry to hear that you are depressed.
true. I am unhappy.
```



The Goals of Explainable Machine Learning

Aims to provide insights into the prediction process of a model.

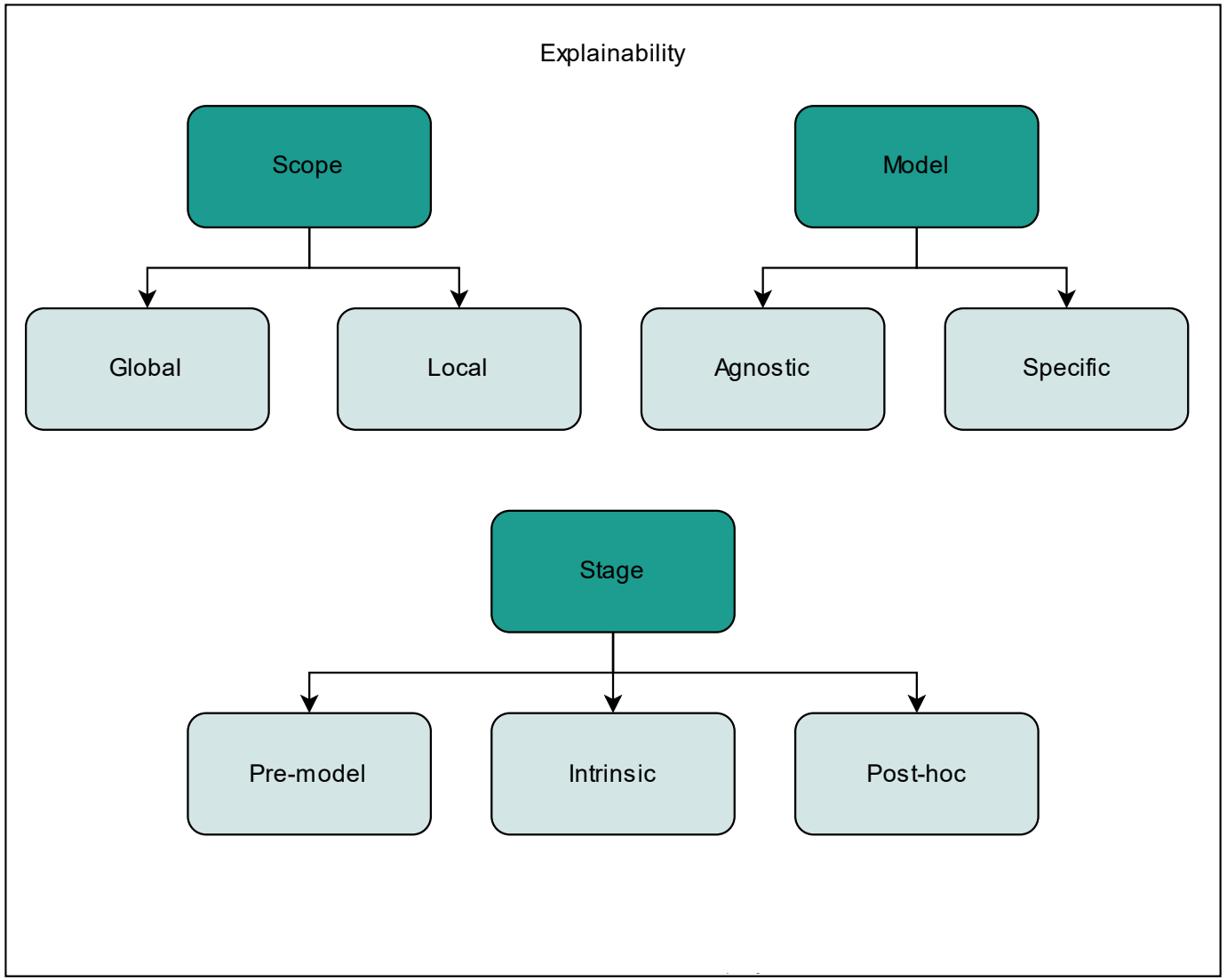
Tool that helps to address numerous challenges to machine learning in the real world

- **Trustworthiness**
- Fairness
- Robustness
- Transparency
- Causality
- Reliability
- ...

Taxonomy of Methods

Explainability methods can be distinguished along 3 axes:

- Stage of application
- Scope of the explanation
- Model specificity.



Intrinsically Interpretable Algorithms vs Blackbox Explanations

Intrinsically interpretable algorithms:

- Self-explanatory. The explanation reflects the true reason for the prediction.
- There is, however, a common belief that interpretability usually comes with a tradeoff in prediction accuracy.

Post-hoc (Black-/greybox) methods

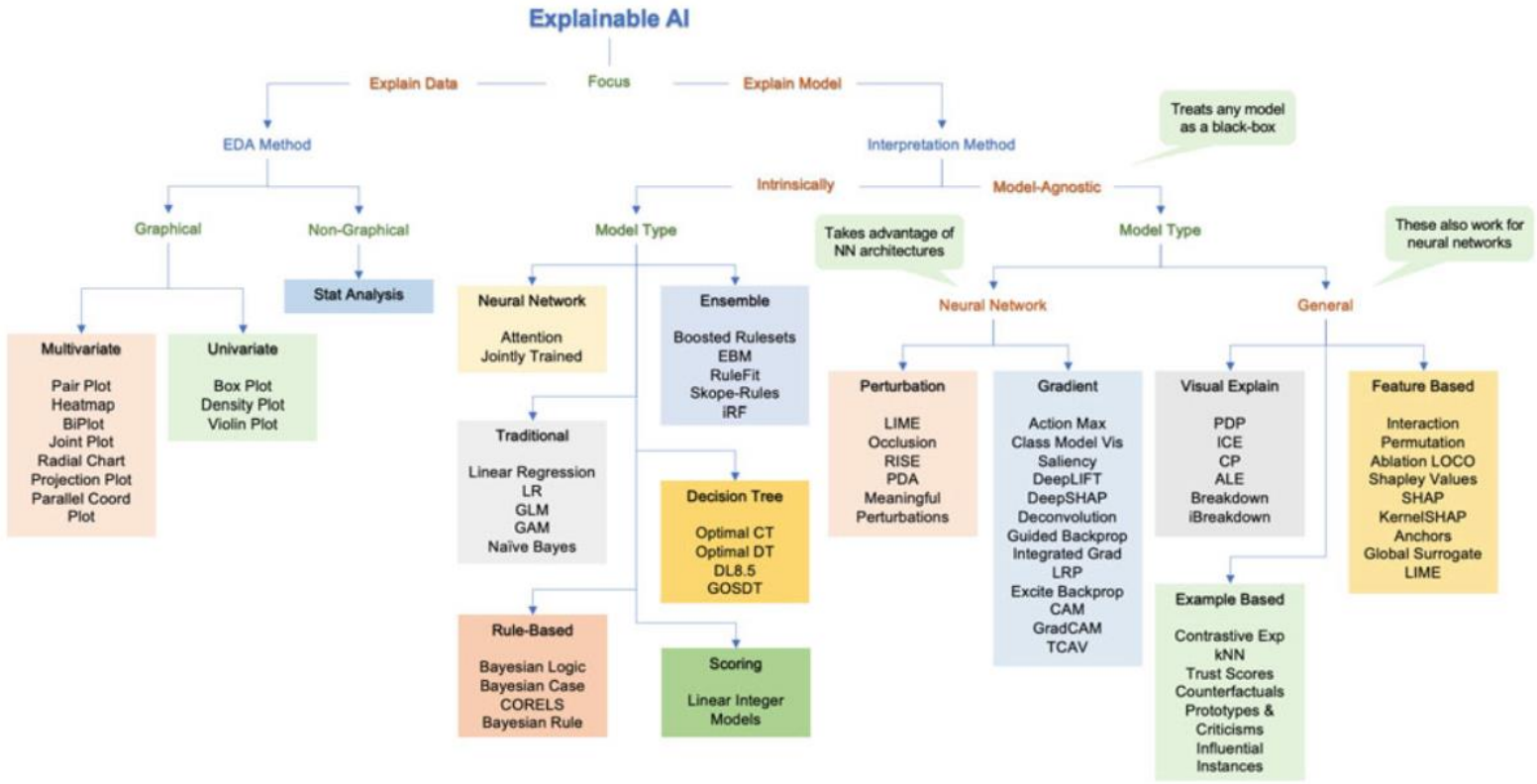
- Attractive because they are applicable to any model, especially including models that are already in use with the need for adoption to the method.
- They are sometimes criticized to confabulate explanations and lend credibility to a decision that might have been taken for completely different reasons.

Types of Explanations

- **Global explanation**: Explain the behavior of the model in general, e.g. model graphs
- **Local explanation**: Explain the behavior of the model on a given instance, e.g. feature attribution
- **Counterfactual explanations**: Describe the smallest change to an instance or model that is necessary to change the outcome into a desired direction.
- **Contrastive explanations**: Describe why the model has chosen X instead of Y. Which properties are more typical for X than for Y?
- **What-if explanation**: Sensitivity analysis of input / model parameters
- **Example-based explanations**: Explanation by examples, e.g. nearest neighbors in the training samples

Algorithm Landscape

- Landscape of Explainable AI algorithms
- Taken from "Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning"



Properties of explanations

Accuracy

- How well does the explanation perform as a predictor.
- Usually a secondary measure. Low accuracy can be fine if the original model has low accuracy.

Fidelity

- How well does the explanation approximate the behavior of the model?
- Essential for most applications.

Comprehensibility

- Is the explanation understandable to humans?
- Similar in importance to fidelity but harder to asses.

Certainty

- Does the explanation quantify the uncertainty in the model prediction.
- Many models provide no or only poorly calibrated uncertainty estimates.

Explainability Along the ML Pipeline

Data analysis: Statistical patterns help to identify important concepts as basis for a meaningful explanations.

Feature **engineering and selection**: Interpretability of features is important for many explainability approaches.

Model selection: Choice of model influences the applicable methods. Interpretability vs accuracy tradeoff.

Evaluation: Provides useful information about strength and weaknesses of a model which helps to understand the behavior.

A Word of Caution

- In many cases explanations are models with weaknesses just as any other model.
- Misleading explanations can lure the user into a false belief of trustworthiness.
- Explainability is tool that can help us to understand models better, but it is not a source of definite truth.

Intro Summary

- Deploying a model that is not well understood in the real world can have unexpected and even severe consequences.
- Explainability aims to provide insights into the prediction process of a model.
- Can be a useful tool to assess the fairness, safety, and reliability of a model (among many other applications).
- Methods can be distinguished by their scope, model specificity and application stage.
- The downstream application of explanations determine the necessary notion of explainability and type of explanation.

Methods and Issues in Explainable AI Workshop Overview

Description

- Introduction to explainable AI
- Reviews many SOTA approaches
- Explores evaluation methods for explanations.
- End-to-end examples
- XAI along the entire ML workflow
- Transferlab page

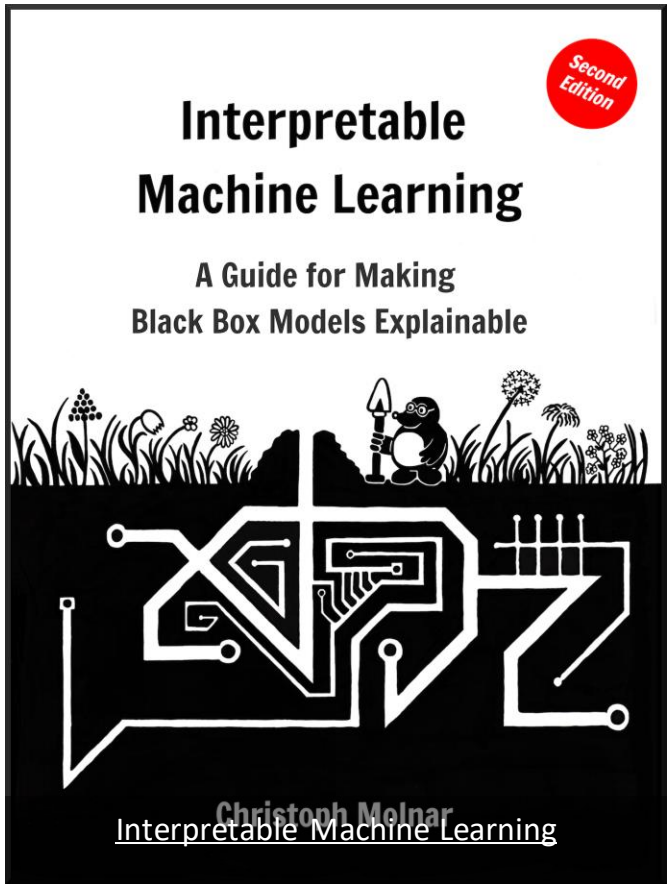
Content

- Post-hoc explainability
  - Partial dependence, accumulated local effects.
  - Lime, Shapley values
- Deep Learning specific methods
  - Data valuation: Influence functions
  - Saliency: Integrated gradients
- Interpretable computer vision
  - Prototype based models
- Interpretable time series forecasting
  - Probabilistic: Prophet
  - Deep: NHIST, Temporal Fusion Transformer

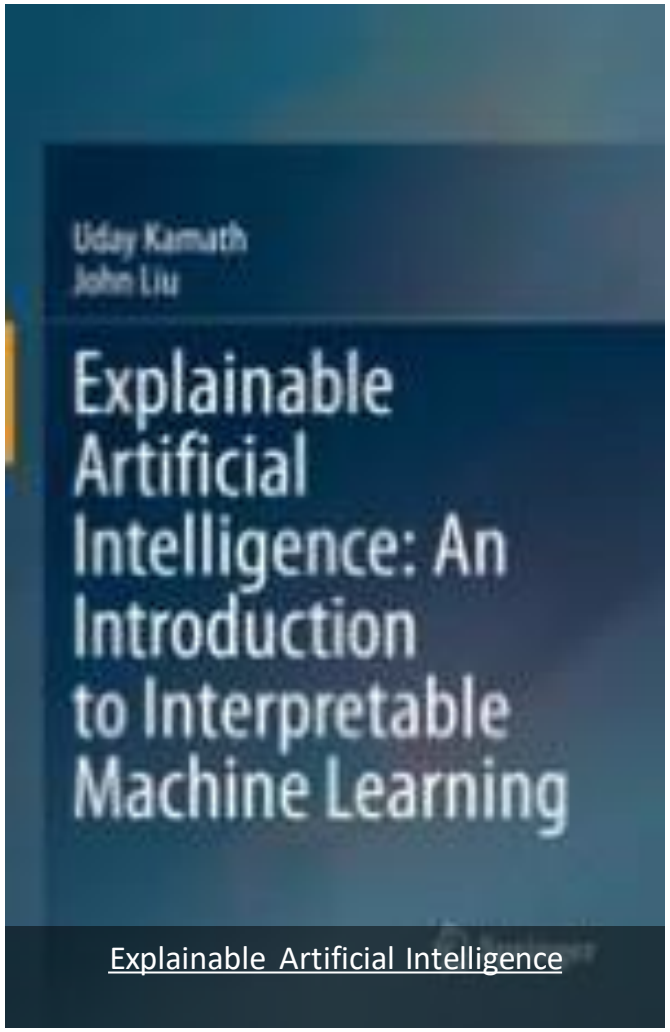
Seminar Topics

1. Introduction to explainable AI
2. The debate on the accuracy-interpretability tradeoff
3. Shapley values for XAI: the good, the bad and the ugly
4. Concept Activation Vectors
5. Counterfactual explanations
6. An information-theoretic perspective on model interpretation
7. Latent space prototype interpretability: Strength and shortcomings
8. Influence functions and Data Pruning: from theory to non-convergence
9. Effects of XAI on perception, trust and acceptance

Books



Interpretable Machine Learning



Explainable Artificial Intelligence

Caution: quite a few errata in 1st edition