Explainable AI (XAI)

September 10th, 2024

Regnskap Norge – OVERSKUDD 2024

Claudia Battistin, PhD

AI Program Director, R&I – Telenor Group

Sometimes AI makes ***small mistakes…***

ChatGPT

Q: *Who is Claudia Battistin?*

A: Claudia Battistin is an Italian academic and professional known for her expertise in international relations, with a particular focus on conflict and post-conflict situations, humanitarian interventions, and the impact of war on societies. She has been involved in various roles related to international cooperation, peacebuilding, and the protection of human rights in conflict zones.

Battistin has also contributed to the academic field through research and publications, particularly on issues related to conflict resolution, peace processes, and international law. Her work often intersects with the efforts of international organizations, NGOs, and other institutions dedicated to addressing the complex challenges of war and peace.

… sometimes makes *big mistakes*

ProPublica

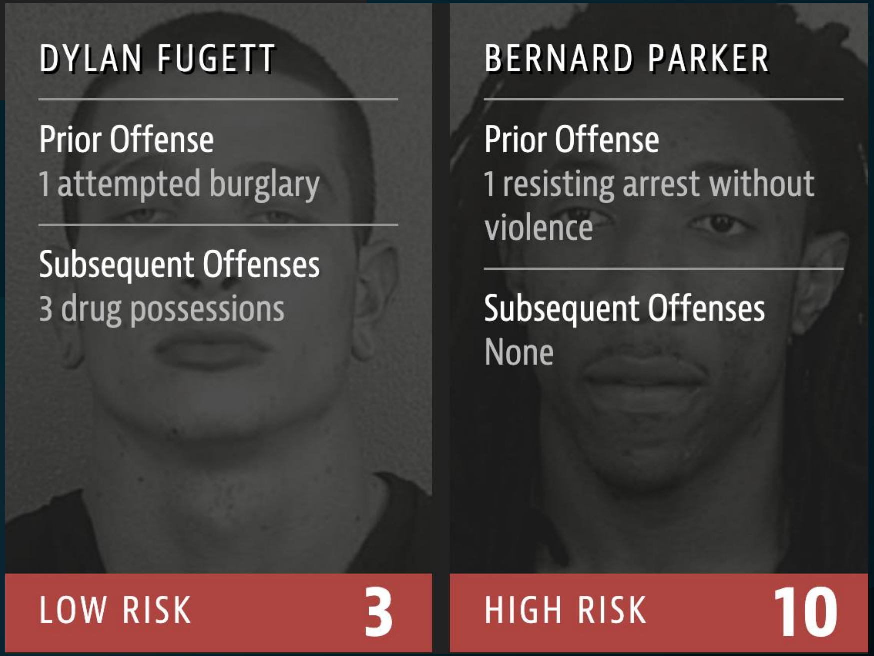
Donate

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016



Why does AI fail?

Training data

- Poor quality (standard, size, ...)
- Skewed
- Not representative of production data

AI design

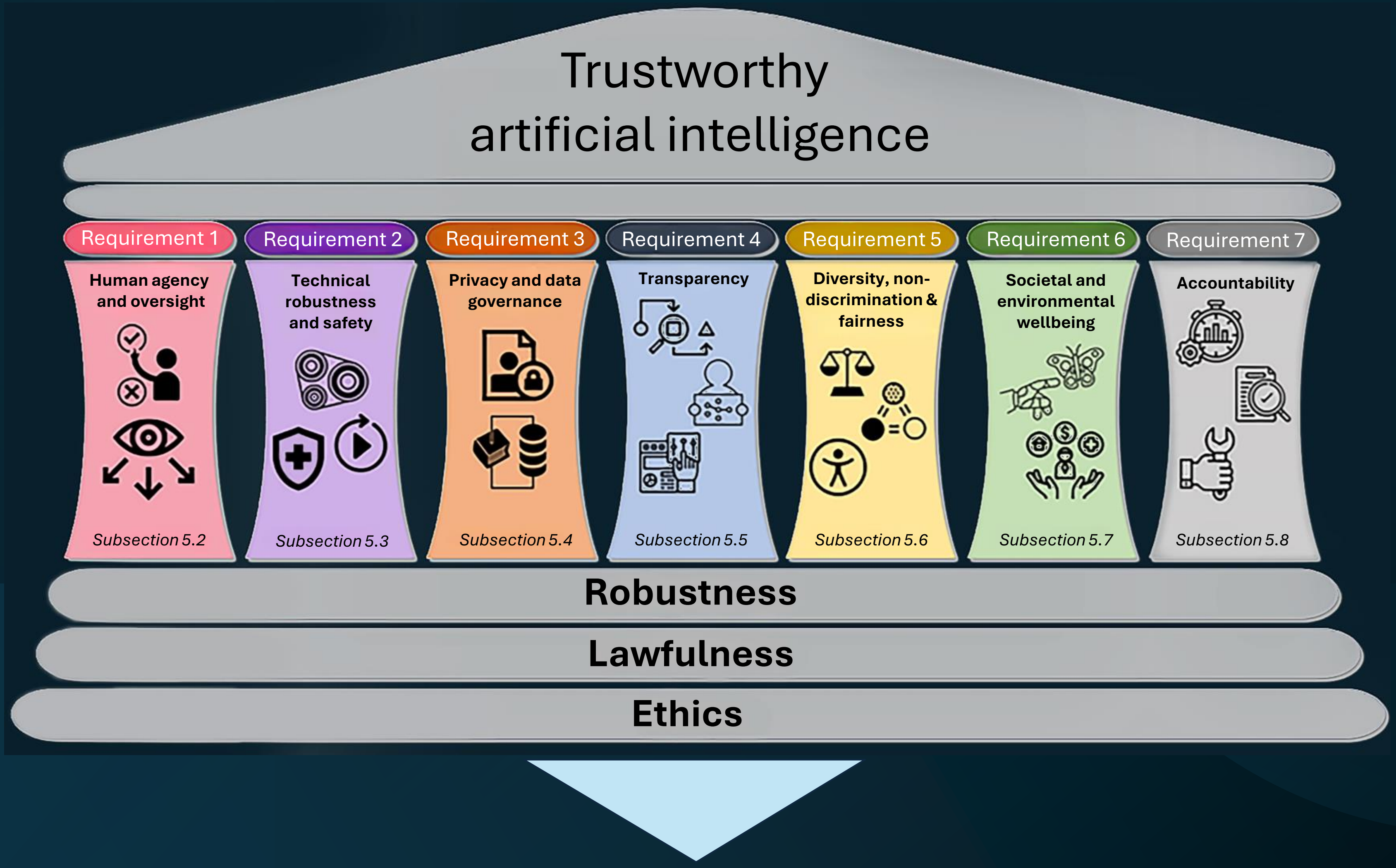
- Poor feature selection
- Spurious correlations

User prompt

- Limited instructions for use
- Absence of guardrails

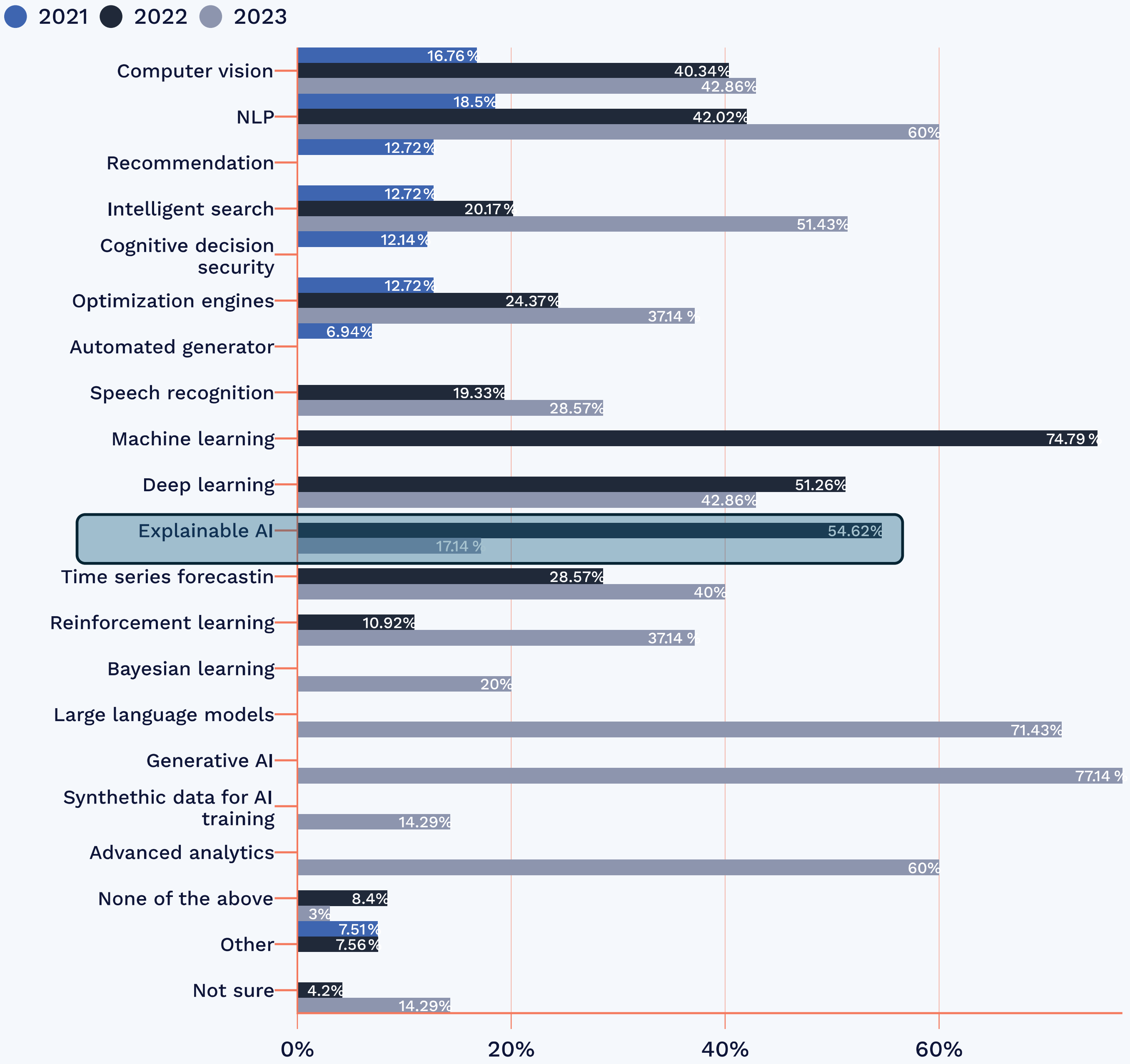
… for high stake decisions

Explainable AI – trust enabler



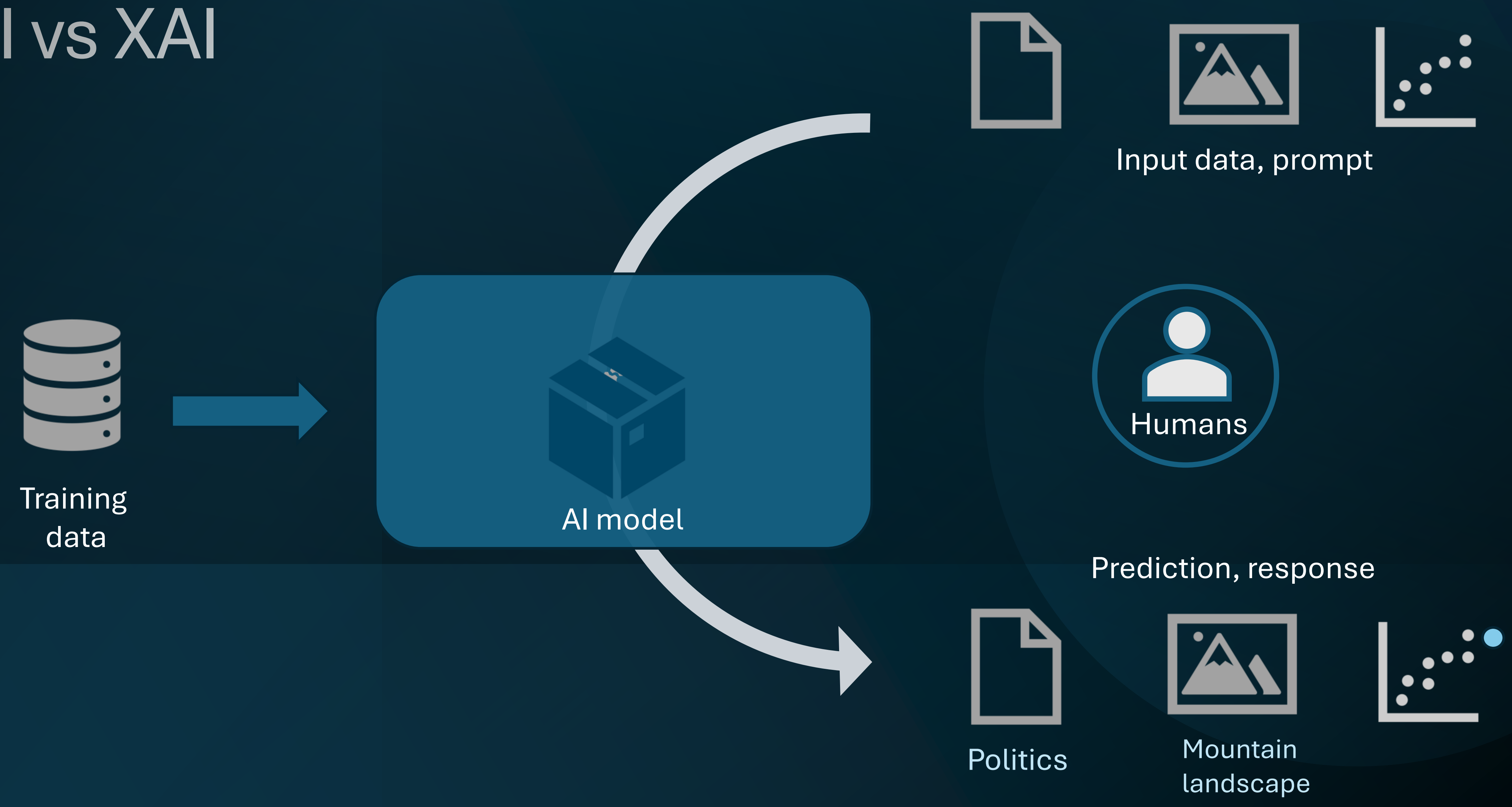
Explaining model and predictions

Explainable AI in the Nordics

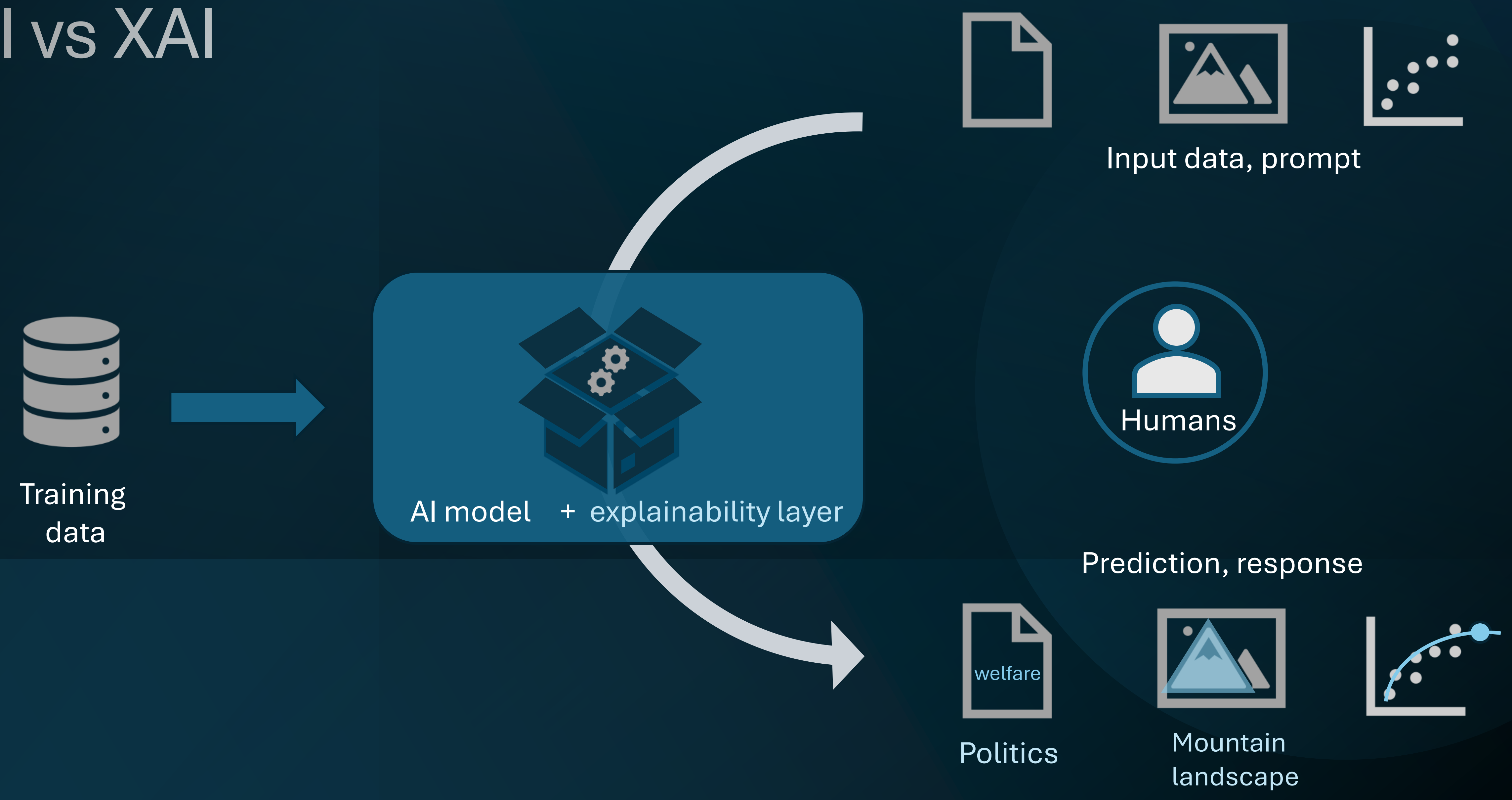


The Nordic State of AI, 2024 – Silo AI

AI vs XAI

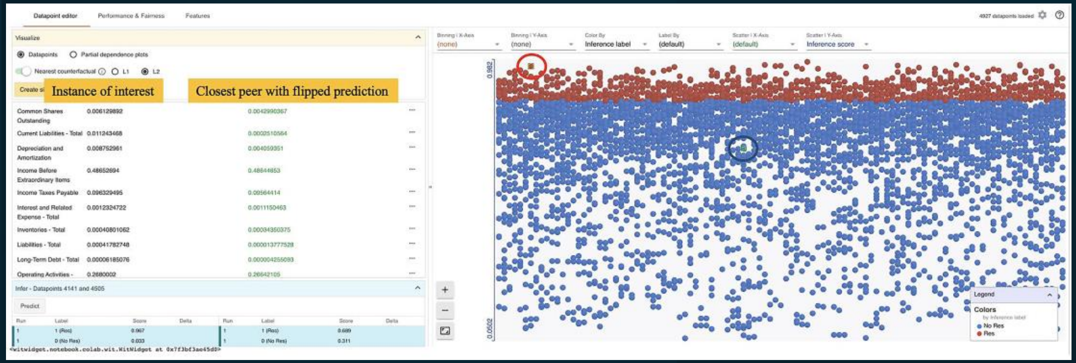

Input data, prompt
Training
data
AI model
Humans
Prediction, response
Politics
Mountain
landscape

AI vs XAI

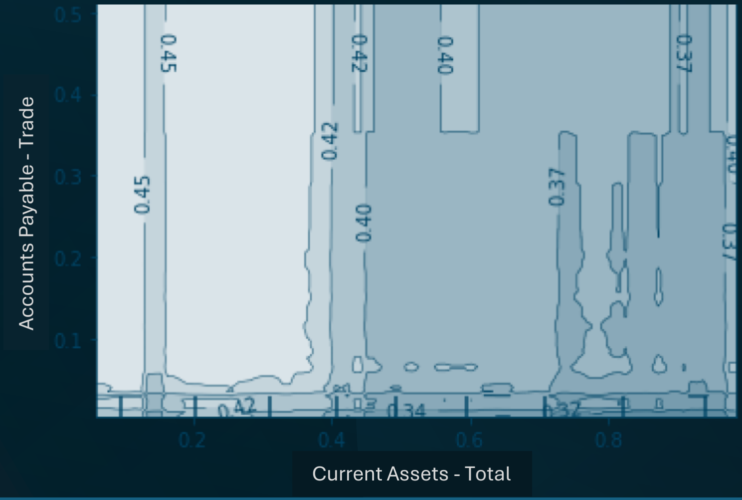

Input data, prompt
Training
data
AI model
+ explainability layer
Humans
Prediction, response
welfare
Politics
Mountain
landscape

Explainable AI approaches

Model agnostic



Feature attribution – Counterfactuals

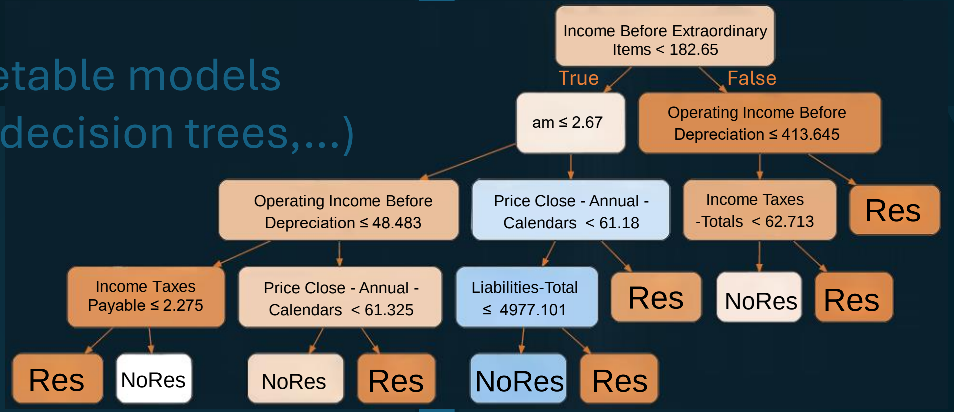


- Surrogate interpretable models
- Stats of local explanations

Local

Global

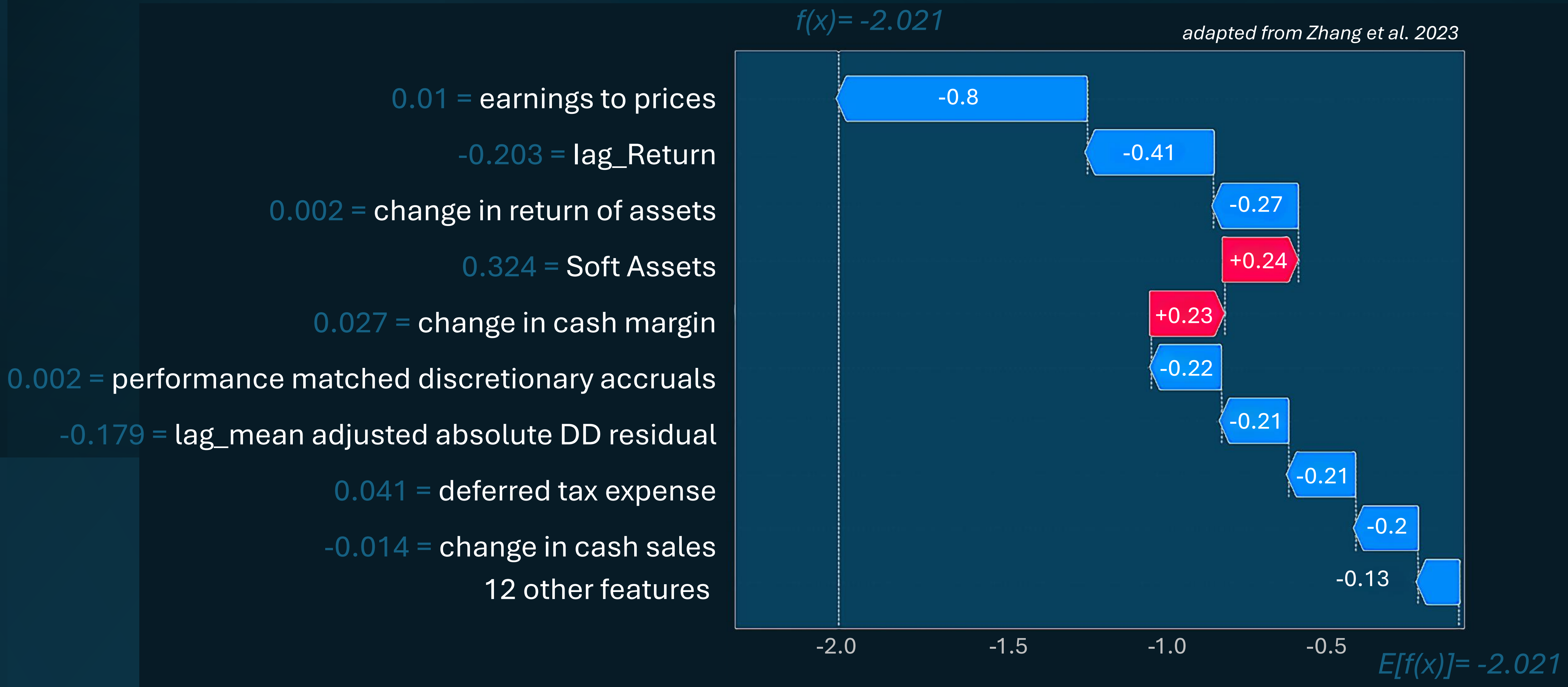
Interpretable models (linear, decision trees,...)



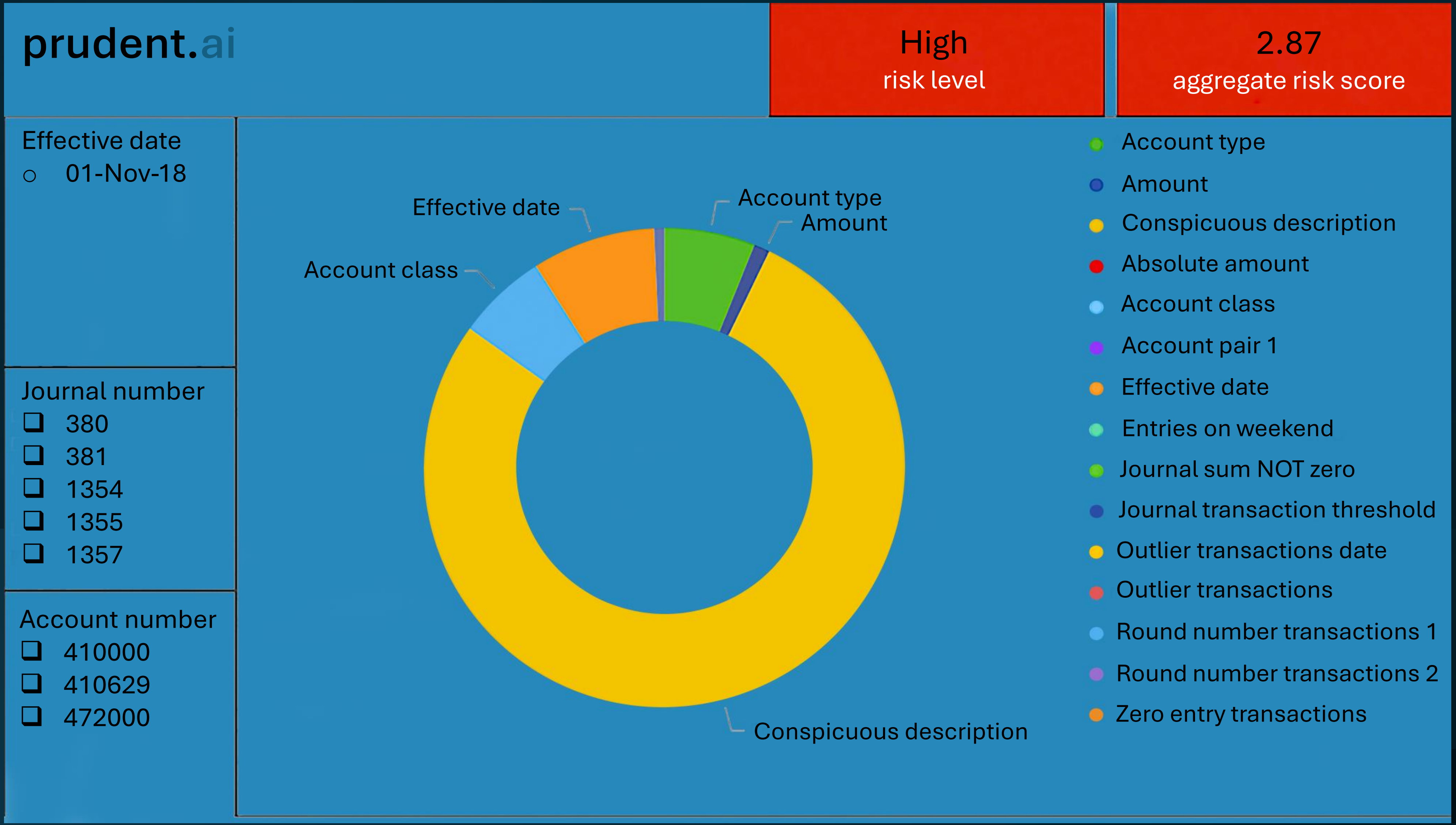
Model specific

adapted from Zhang et al. 2023

Probability of material restatement

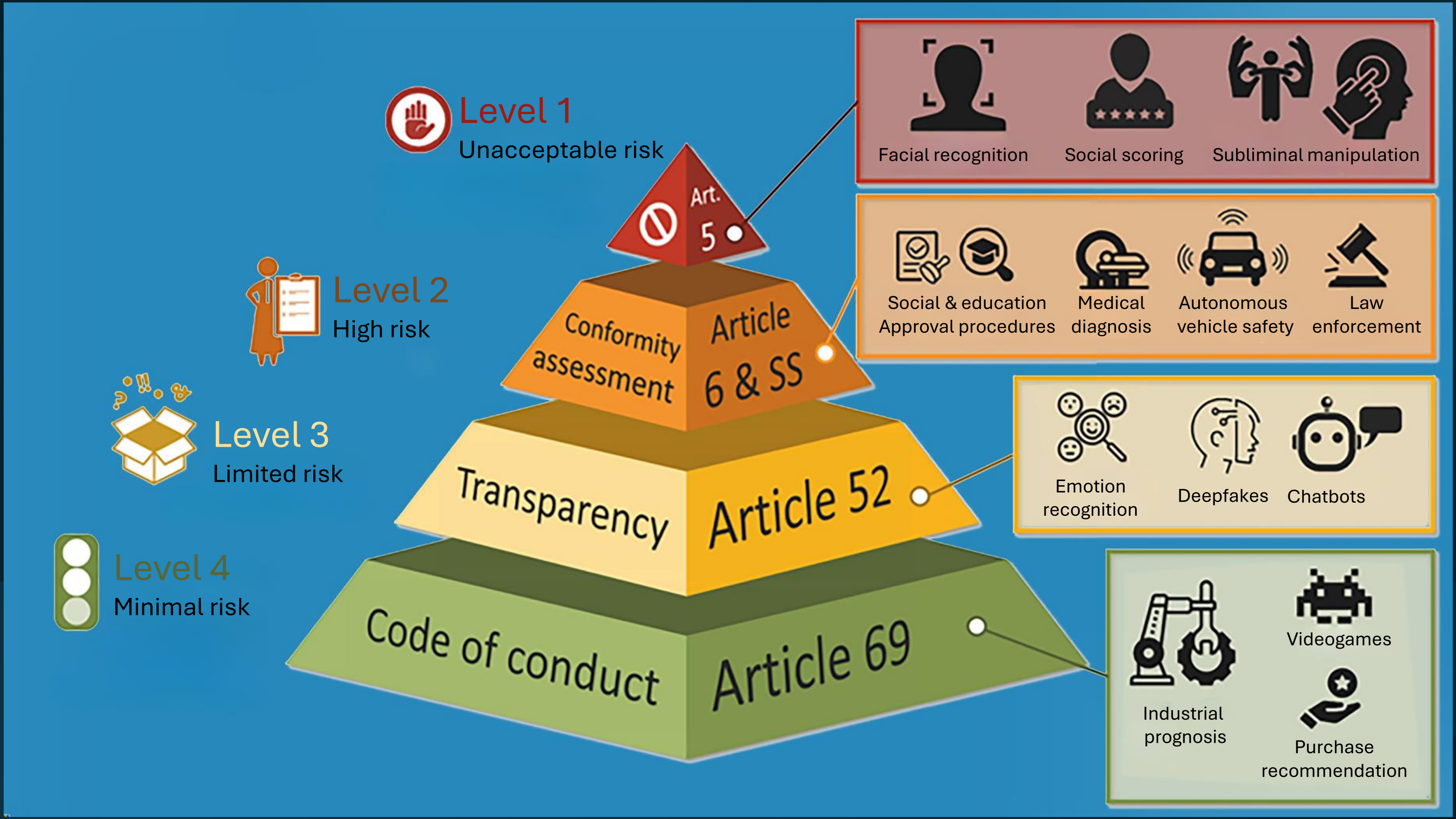


Risk of transaction



adapted from Vaidyanathan, ACCA

Explainable AI – EU AI Act



adapted from Díaz-Rodríguez et al. 2023

Transparency
- Documentation
- Instruction to users (scope, limitations, error rate, GLOBAL EXPLANATIONS, ...)
- Watermarking

Human oversight
(LOCAL EXPLANATIONS?)

Explainable AI – GDPR

Right to explanation

<u>Recital 71 (not binding)</u> – LOCAL EXPLANATIONS

The data subject should have the right […] to obtain an explanation of the decision [solely on automated processing] reached after such assessment and to challenge the decision.

<u>Article 13(2)(f)</u> – GLOBAL EXPLANATIONS

In the event a controller relies on automated decision-making, […] the controller is under the duty to provide the data subject with "meaningful information about the logic involved" […] of such forms of processing.

Why explainable AI in accountancy?

Understand AI predictions

Professional skepticism

Foster AI Innovation

Thanks