Contradictio in Terminus: Explainable AI?

Prof. Dr. Steef Peters /
Dr. Gerrit Jan van den Brink

The Hague University of Applied Sciences

EAA e-Conference on
Data Science & Data Ethics

14 May 2024

- Most recent developments:

  - Developments in the legal area

  - Regulatory developments around the "AI explainable" requirement

  - Use of norms for AI applications

  - Latest scientific requirements around explainable AI

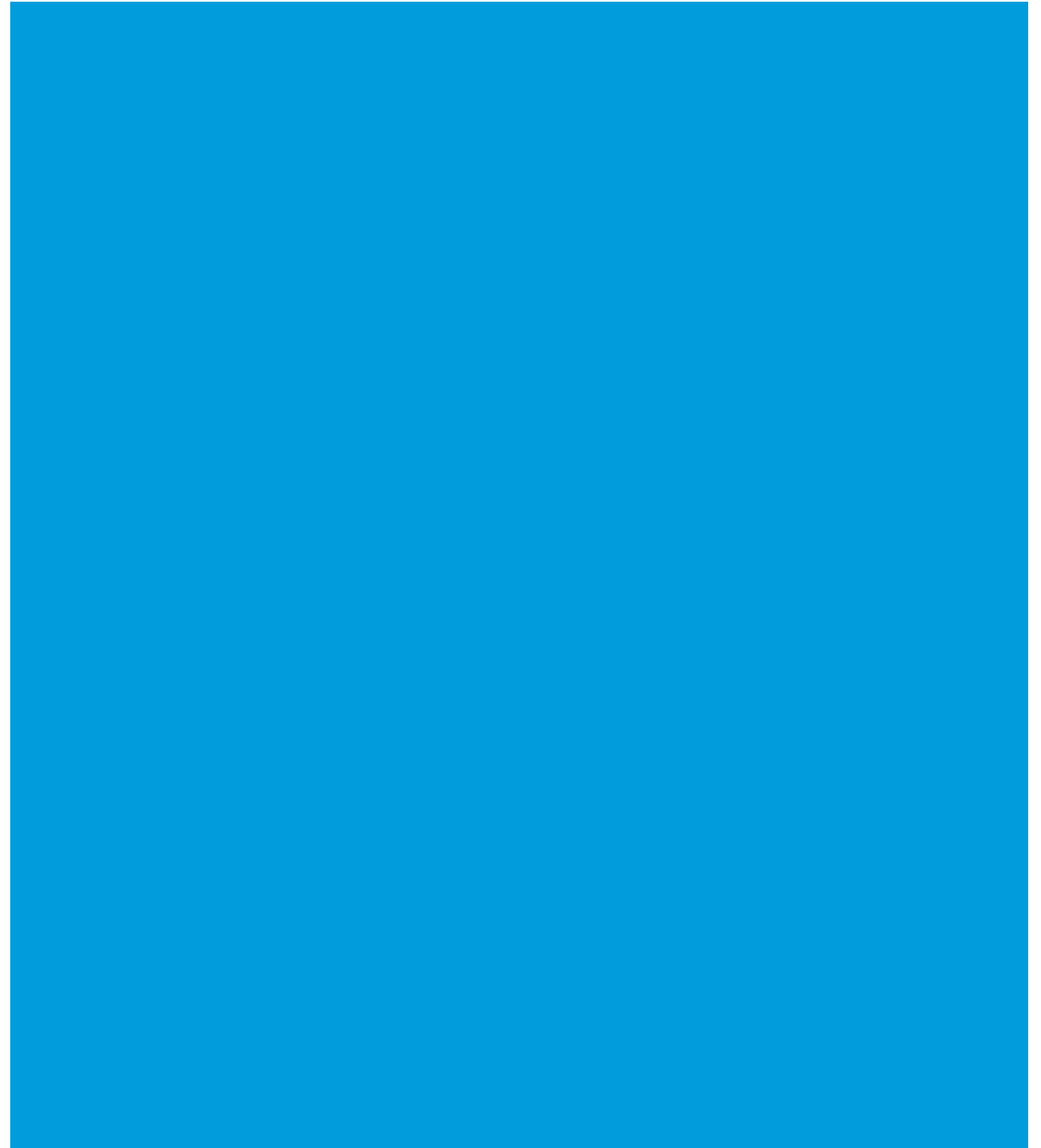- Can AI have a moral and if so how would it work and be governed?
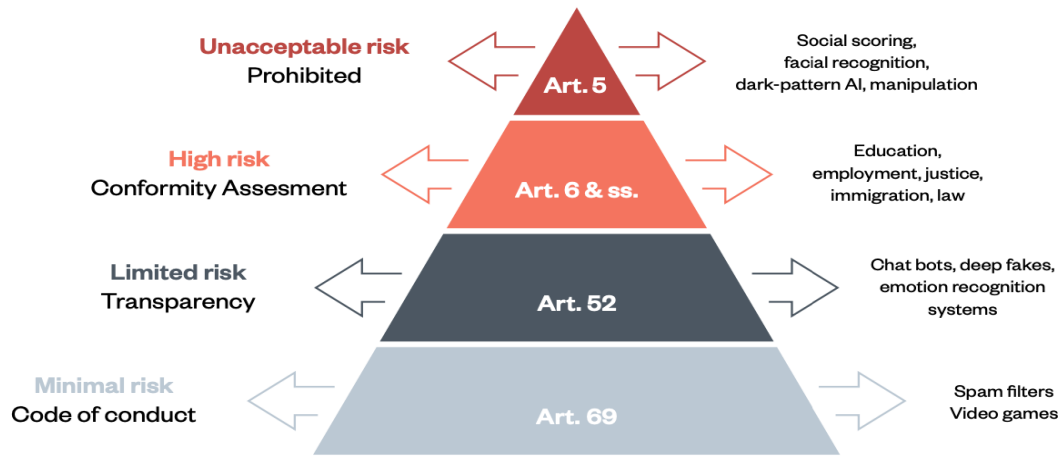
- Next steps

MOST RECENT DEVELOPEMENTS

EU HAS MOST ADVANCED AI GOVERNANCE FRAMEWORK

Source: L. Edwards, The EU AI Act: a summary of its significance and scope, 2022, p. 9

AI Safety: UK and US sign landmark agreement

1 day ago

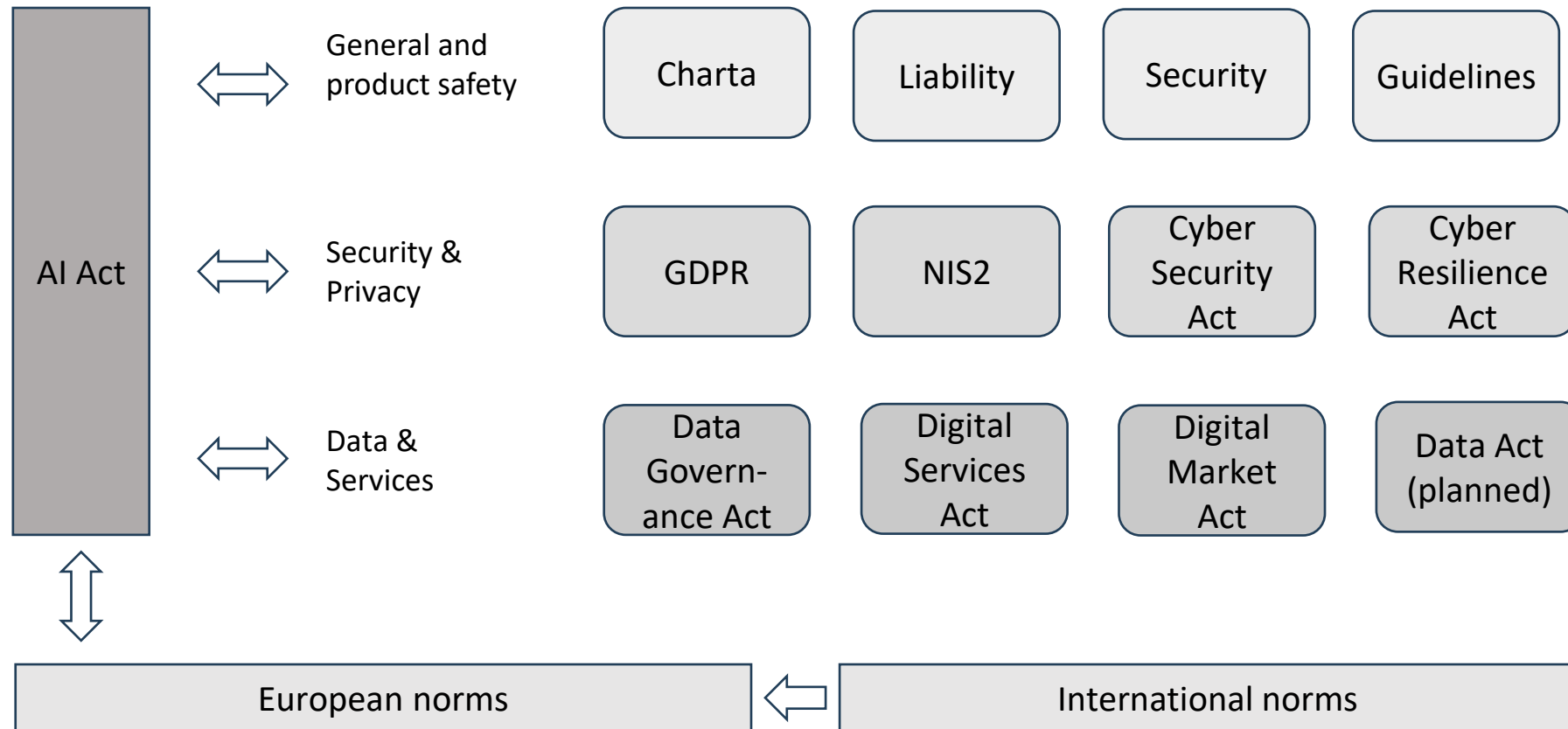By Liv McMahon & Zoe Kleinman, BBC News

- The EU has taken a conservative legal view on AI governance by keeping the makers (Natural persons/legal entities) fully accountable for the AI product and its use

- The UK and US seem to take a slightly more accommodating approach:

  - Less strict disclosure requirements

  - Balanced focus on opportunities

OVERVIEW WITH EU-LAWS RELATED TO THE AI ACT

```
                    General and      ┌─────────┐  ┌─────────┐  ┌─────────┐  ┌─────────┐
                    product safety    │ Charta  │  │ Liability│  │ Security│  │Guidelines│
          ⟷                           └─────────┘  └─────────┘  └─────────┘  └─────────┘

┌────────┐                           ┌─────────┐  ┌─────────┐  ┌─────────┐  ┌─────────┐
│        │          Security &        │  GDPR   │  │  NIS2   │  │  Cyber  │  │  Cyber  │
│ AI Act │   ⟷      Privacy          │         │  │         │  │ Security│  │Resilience│
│        │                           └─────────┘  └─────────┘  │   Act   │  │   Act   │
│        │                                                     └─────────┘  └─────────┘
│        │          Data &           ┌─────────┐  ┌─────────┐  ┌─────────┐  ┌─────────┐
│        │   ⟷      Services         │  Data   │  │ Digital │  │ Digital │  │Data Act │
└────────┘                           │Govern-  │  │ Services│  │ Market  │  │(planned)│
                                     │ ance Act│  │   Act   │  │   Act   │  │         │
                                     └─────────┘  └─────────┘  └─────────┘  └─────────┘
    ⟷

┌──────────────────────────────┐              ┌──────────────────────────────┐
│      European norms          │     ⟵        │     International norms       │
└──────────────────────────────┘              └──────────────────────────────┘
```

EUROPEAN HIGH COURT DECISION

ECJ on the compatibility of scoring with the provisions of the GDPR (C-634/21)

The ECJ ruled that scoring constitutes an "*automated individual decision-making*", which is generally not permitted under the GDPR, if the scoring determines whether a third party to whom the score value is transmitted establishes or terminates a contractual relationship with this person. The referring court must examine whether the BDSG contains an effective exception to this prohibition and, if this is the case, whether the general requirements of the GDPR for the processing of personal data are met.

According to the ECJ, scoring constitutes such an automated decision-making if the customers of SCHUFA, e. g. banks, make their decision (e. g. on the granting of credit) solely dependent on the score value.

Aus Sicht von Bundesverbraucherschutzministerin Steffi Lemke (Grüne) hat der EuGH die Verbraucherrechte beim Scoring gestärkt. „Mit dem Urteil wird der Schutz der Verbraucherinnen und Verbraucher erweitert: Wer einen Vertrag abschließen will, muss sich darauf verlassen können, dass dieser nicht maßgeblich durch eine Maschine abgelehnt wird", sagte sie.

- Black box algorithms only are not allowed for credit decision making
- Human interference in high impact decisions is required
- German Federal Minister responsible for consumer rights stated clearly that consumers can rely upon the fact that rejections are not only based on machine decisions
- Based on discussions the human interference is only needed in case of rejections. Approvals can be used directly

SUPERVISORS SHARED THEIR VIEWS: EXAMPLE DUTCH CENTRAL BANK

General principles for the use of Artificial Intelligence in the financial sector

DeNederlandscheBank

EUROSYSTEEM

Perspectives on Explainable AI in The Financial Sector

- DNB introduced the „SAFEST" principles:
  - Soundness
  - Accountability
  - Fairness
  - Ethics
  - Skills
  - Transparency (now often referred to as explainability)

- A study with participation of banks, supervisors and the University of Applied sciences in Utrecht showed a divergence in opinion around the explainability requirement
- Discussion needs to be intensified to prevent implementation hurdles

SUPERVISORY CONCERN MAINLY IN THREE AREAS
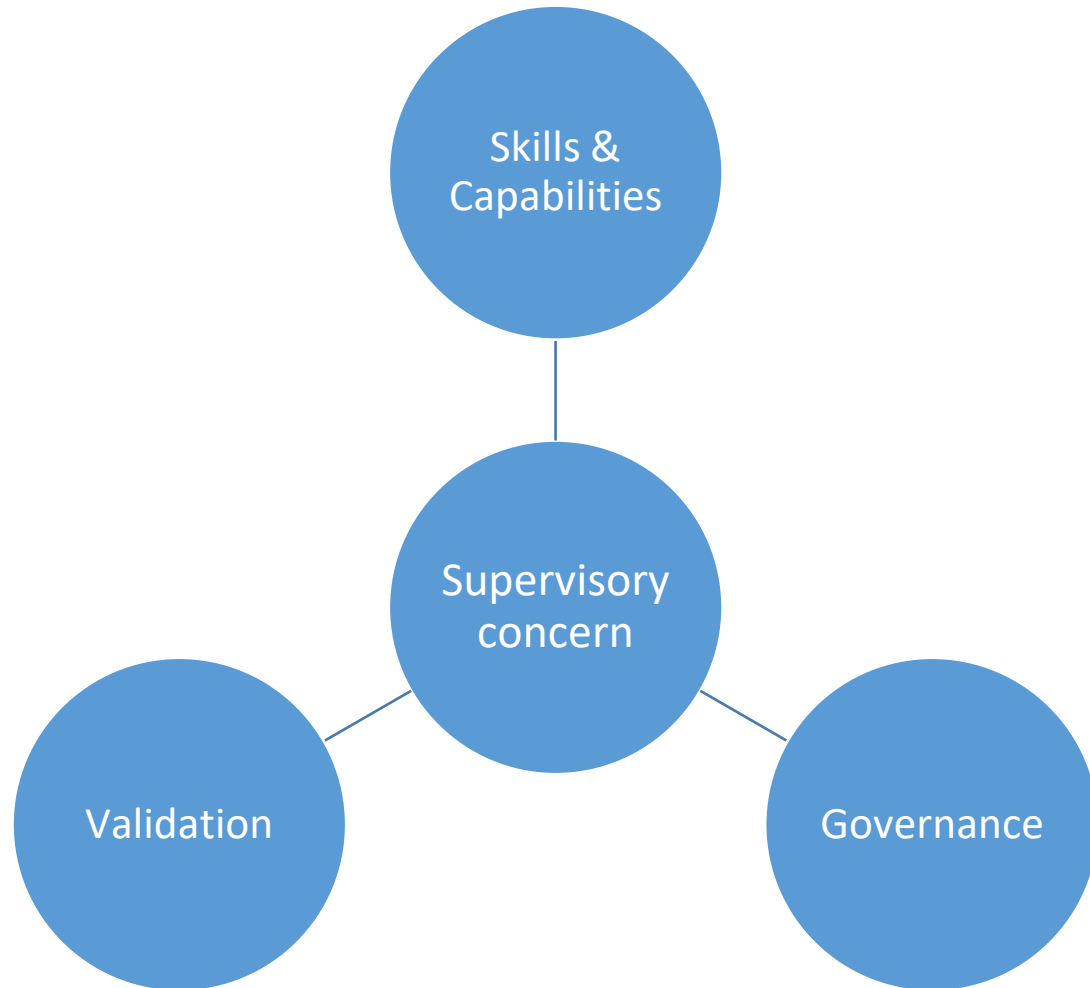
- Skills and capabilities for both the users and creators of models

- Are supervised organisations in full control of all model components (e.g. foundation models)?

- AI models should be considered like all other models and therefore be subject to the same model governance rigor as other (internal) models

- This includes use test components comparable to Solvency II models

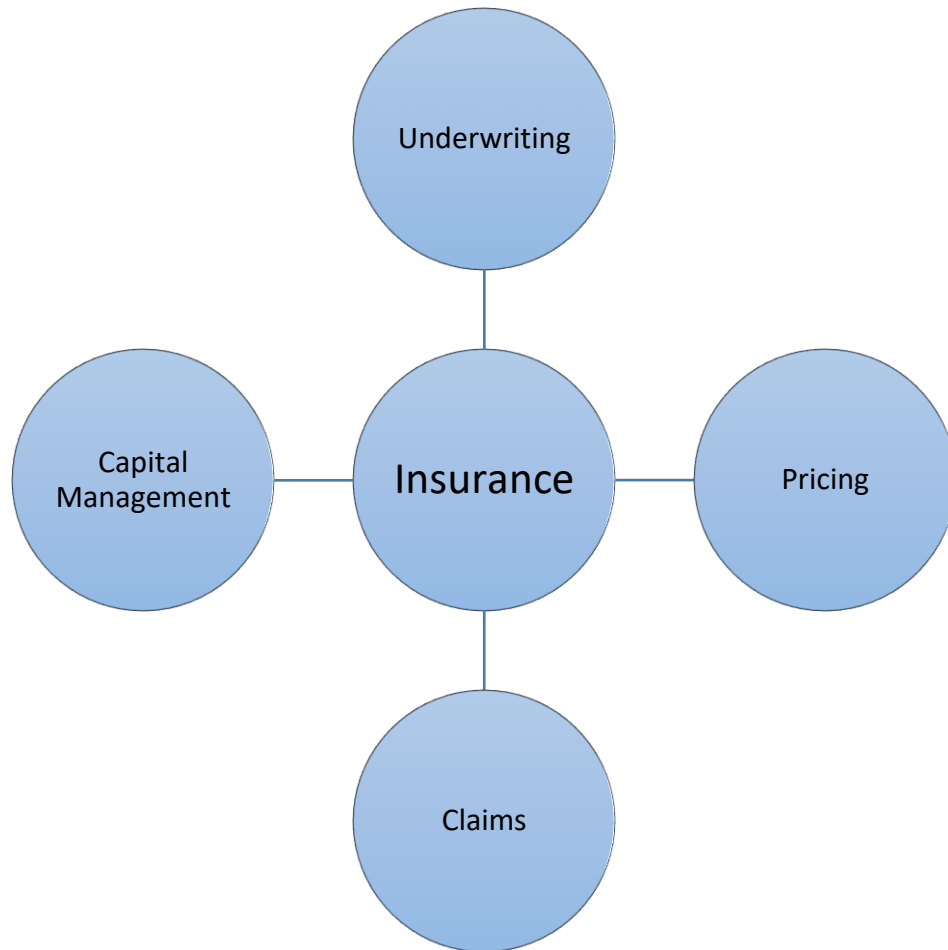# A MULTITUDE OF NORMS HAS BEEN DEVELOPED

Essential components for trusted AI

- Legitimacy
- Ethical
- Robustness
- On the fundament of
  - Human autonomy
  - Damage prevention
  - Fairness
  - Transparent

Core requirements

- Prevalence of human actions and oversight
- Technical robustness and security
- Data protection and data quality management
- Transparency
- Diversity, non-discrimination and fairness
- Societal and ecological wellness
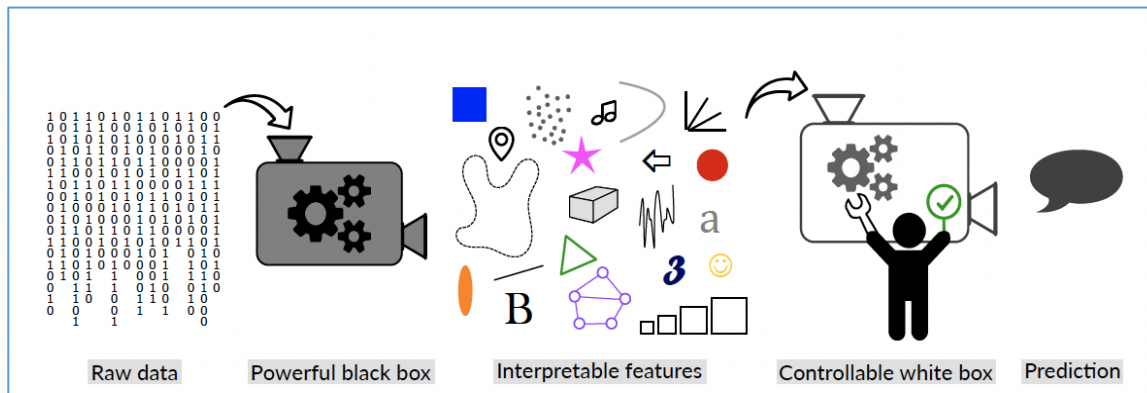- Accountability

Source: DIN, DKE (2022): Deutsche Normungsroadmap Künstliche Intelligenz (Ausgabe 2), 2022, p21

IMPACTED AREAS

Underwriting

Capital Management

Insurance

Pricing

Claims

Areas to pay attention to:

- Vulnerable customers

- Potential discrimination due to:
  - Biased data sets
  - Model design

- Hallucination

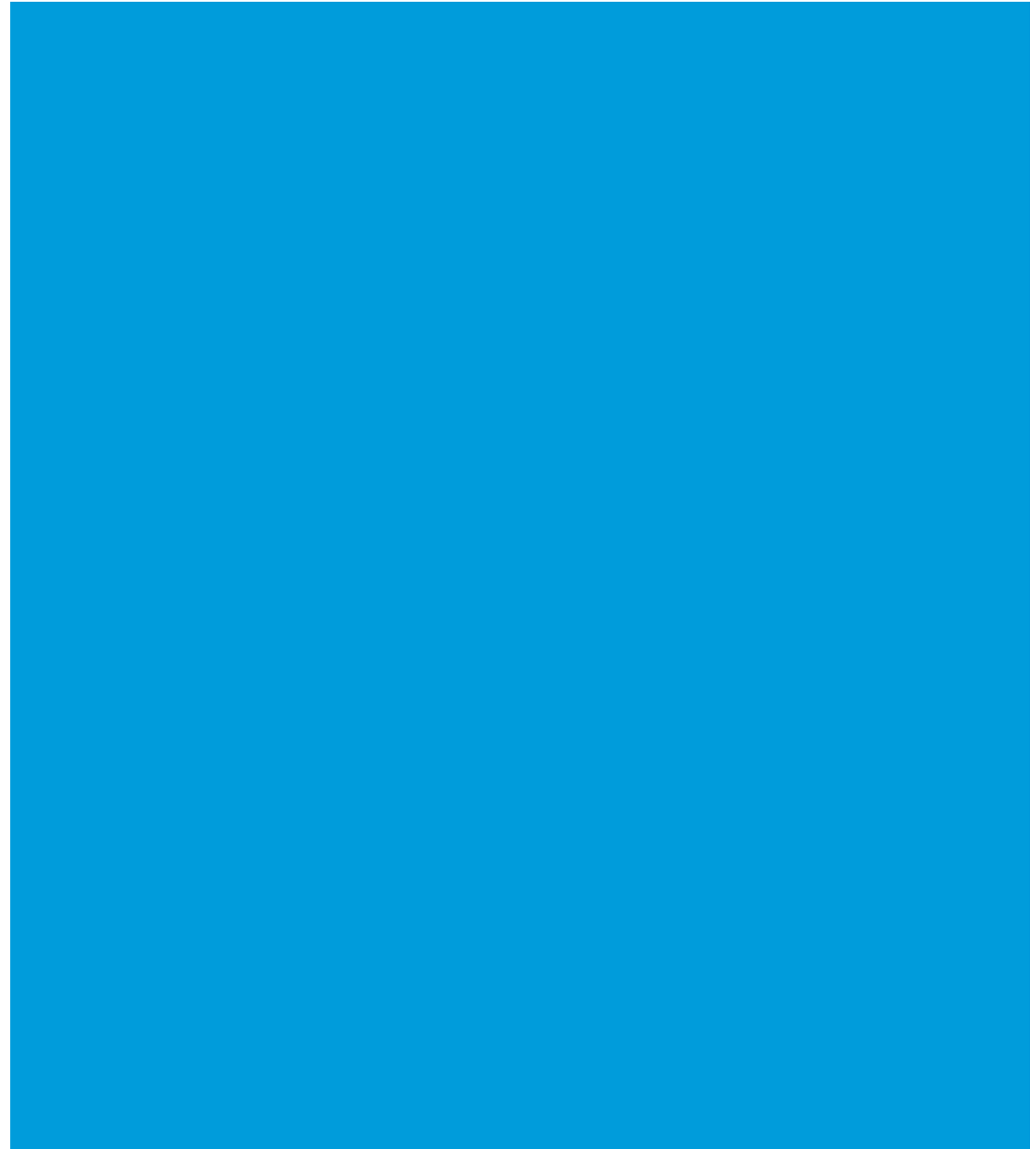- Explainability of decisions in all (esp. customer related) areas

MEIKE NAUTA MADE GREAT EFFORT ON MAKING AI EXPLAINABLE

EXPLAINABLE AI AND INTERPRETABLE COMPUTER VISION

from oversight → to insight

MEIKE NAUTA

Raw data — Powerful black box — Interpretable features — Controllable white box — Prediction

Co-12 property		Evaluation approaches (extract)
		CONTENT
$f \stackrel{?}{=} e$	Correctness	• Classification process of part-prototype models is correct by design since $f(x) = e(x)$ • Evaluate prototype visualisation with synthetic data or incremental deletion/addition of image patches
	Completeness	• Output-complete by design • Evaluate human-output-completeness with simulatibility user studies
$e = e$	Consistency	• Implementation invariance and nondeterminism
$e \approx e$	Continuity	• Stability for slight variations
$e \leftrightarrow e'$	Contrastivity	• Contrastive by design; can answer counterfactual questions • Pragmatism and compactness for optimal contrastive explanation • Target-sensitivity for location of prototypes • Target-discriminativeness to evaluate prototypes

- XAI can provide answers to the question: Why did the model predict this?

- XAI can provide insight in undesired patterns in the data: right predictions for the wrong reasons

- Split modelling in a predictive and explanation-centric path

- Models should be trained to learn task-relevant interpretable features

- CO-12 properties can define explainability quality

CAN AI USE MORALITY AND IF SO HOW WOULD IT WORK AND BE GOVERNED?

In the use of AI algorithms we can distinguish three levels:

1. Self learning general AI algorithms based on external data

2. Self learning specialized AI algorithms based on specific data offered by experts

3. Governance AI algorithms based on specific data regarding norms and rules (laws)

Governance

Expert

General
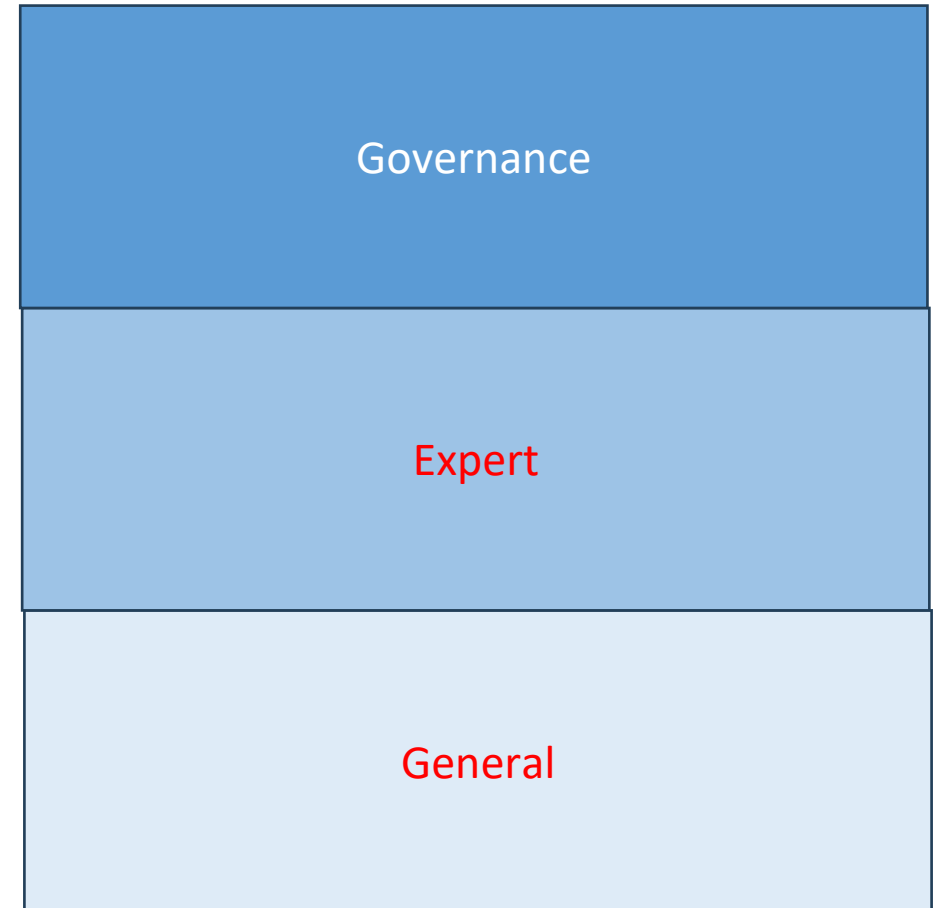
- Expert AI algorithm uses results of earlier decisions of experts in comparable situations to determine „most wanted" result

  When needed AI algorithm asks for input of expert

- Governance AI algorithm uses existing data containing existing rules and norms destilled from earlier decisions to govern the action to get "acceptable results"

  When needed the AI algorithm asks for input from monitoring expert (for instance compliance manager)

  Symbolic language necessary for describing laws, rules and norms in the context of operation

- Three phases for AI algorithms

  - Ex-ante
    Measures to be taken before executing algorithm

  - Processing
    Governance (monitoring) while executing algorithm

  - Ex-post
    Measuring result of AI algorithm andusing it for (later) decisions

Ex-ante | Processing | Ex-post

goal

fulfilment

Organization

Ex ante

rules

Monitoring

result

Ex post

action

anomaly

AI process

input

output
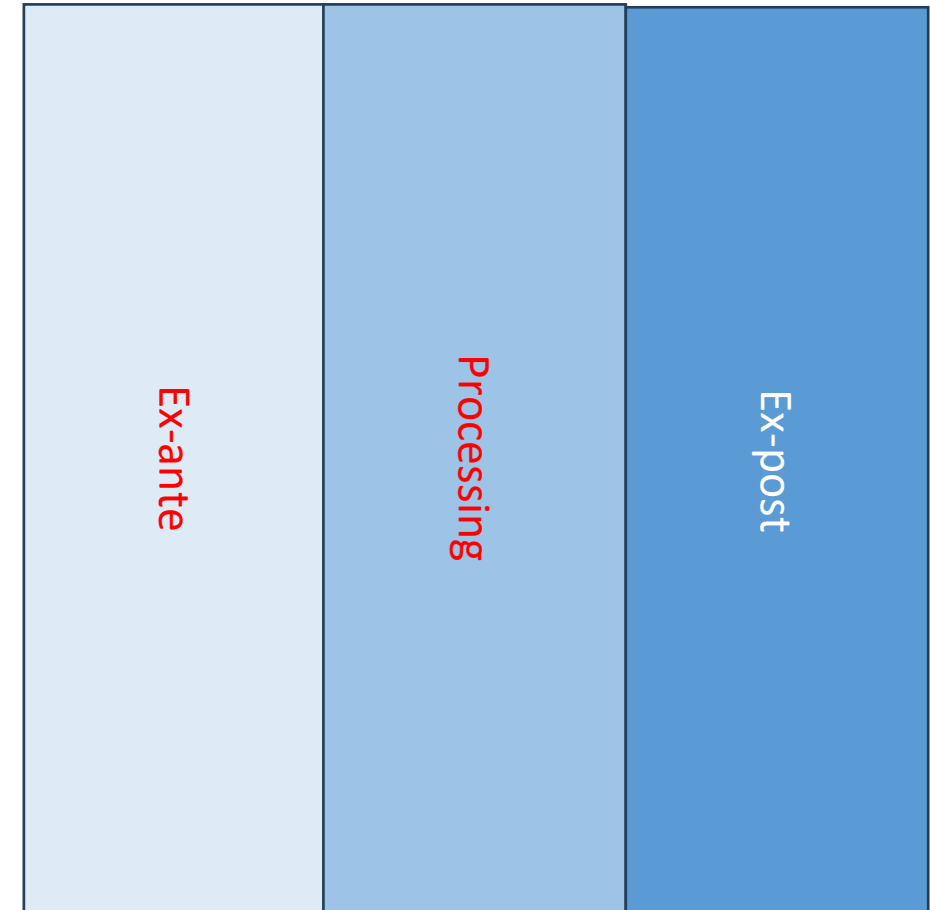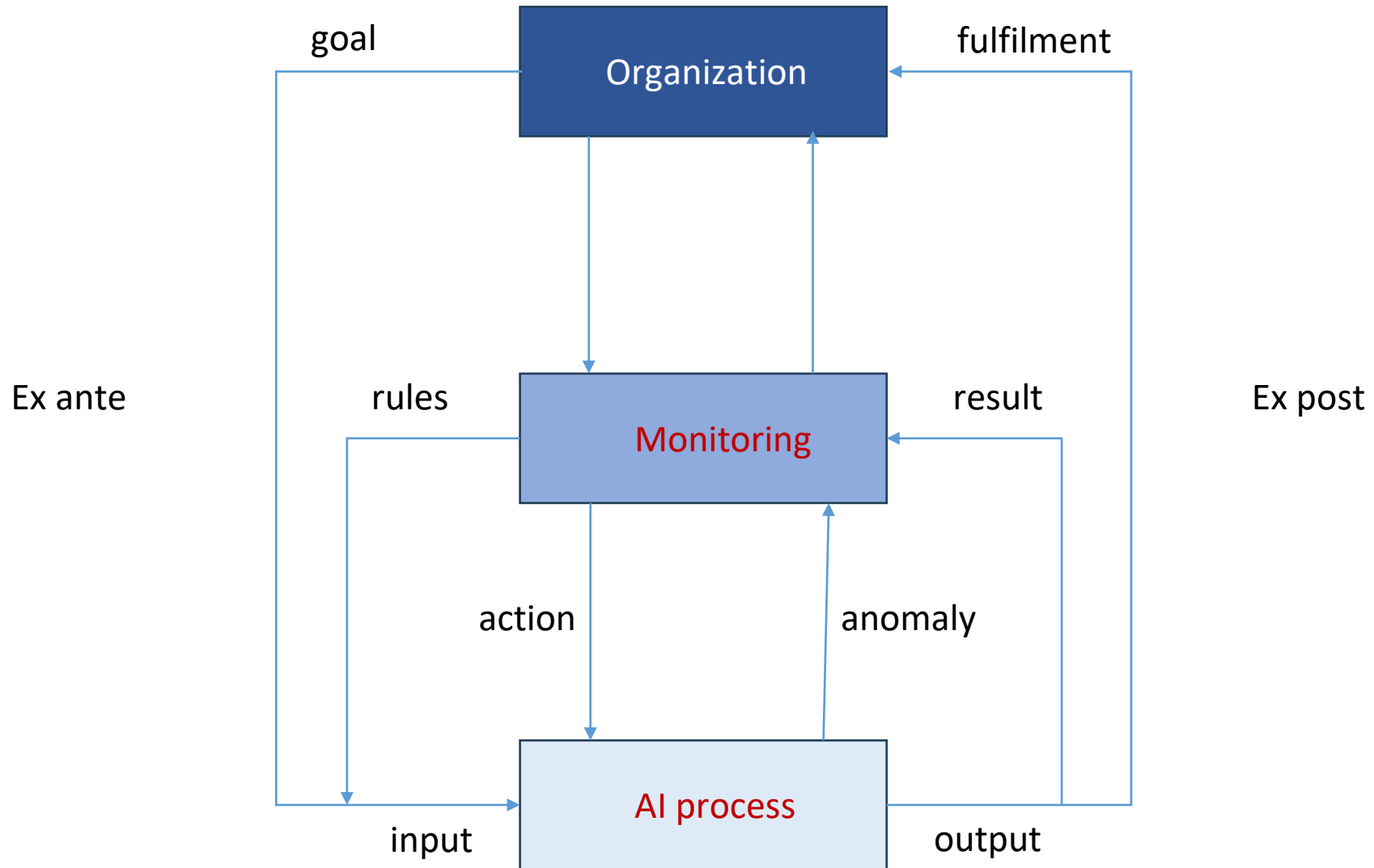
- When using smart contracts, the execution follows same path:

  - Ex-ante determining execution rules smart contract based on available data
  - Execution of smart contract governed by intermediate checks on expert and norm data
  - Ex-post measuring result of smart contract execution and possible acceptance

Steef Peters

- PhD in elementary particle physics
- CIO leasing company
- Startups in DeFi, Global Outsourcing
- Professor in Strategy and Business Informatics

Gerrit Jan van den Brink

- PhD in economics, Chartered Accountant
- CRO insurance company
- Startups in modern consulting businesses
- Lecturer on Risk Management, Cyber Security, ESG, Inflation

ABOUT US

Entrepreneurial experimentalist / wallet79

Founder and MD Risk Sigma GmbH

REFERENCES (EXAMPLES)

Cihon, P., Schuett, J., & Baum, S. B. (2021, July 5). *Corporate Governance of Artificial Intelligence in the Public Interest.* Retrieved September 2022, from Information 2021: https:/doi.rog/10.3390/info12070275

Evrim, T., Stanislav, M., & Joep, C. (2022). *Blockchain governance in the public sector: A conceptual framework for public management.* Retrieved 8 2022, from www.sciencedirect.com: https://doi.org/10.1016/j.giq.2021.101625

Hilb, M. (2020, July 21). Toward artificial governance? The role of artificial intelligence in shaping the future of corporate governance. *Journal of Management and Governance*, pp. 851-870.

Huber, W. (2022). *Menschen, Götter und Maschinen, Eine Ethik der Digitalisierung.* München: C.H. Beck.

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022, Feb. 24). *Defining organisational AI governance.* Retrieved September 2022, from AI and Ethics: https://doi.org/10.1007/s43681-022-00143-x

Peters, S., & Stephenson, K. (2022). *Toward a General Theory of Organizing, Volume 1: Introducing the Network Field Model.* Retrieved from http://dx.doi.org/10.5772/intechopen.99709

Renda, A. (2019). *Artificial Intelligence. Ethics, governance and policy challenges.* Brussels: Center for European Policy Studies.

Taeihagh, A. (2021). *Governance of artificial intelligence.* Retrieved September 2022, from https://doi.org/10.1080/14494035.2021.1928377

Thank you very much for your attention

Contact

Dr. Gerrit Jan van den Brink

The Hague University of Applied Sciences/Risk Sigma GmbH

gjvandenbrink@outlook.de

+49.176.5378 8522

EAA e-Conference on
Data Science & Data Ethics

14 May 2024