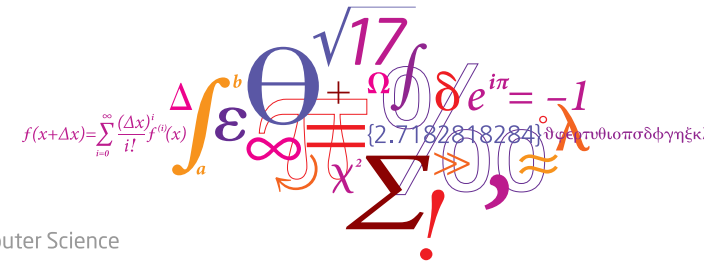
Why 'Trustworthiness' in AI is difficult?

Challenges in Defining, Operationalizing and Regulating Trust

Sneha Das

Assistant Professor (Tenure track)
Dept. of Applied Mathematics and Computer Science
Technical University of Denmark

sned@dtu.dk
https://snehadas.github.io



DTU Compute
Department of Applied Mathematics and Computer Science

Why 'Trustworthiness'?

Why 'Trustworthiness'?

Official Journal
of the European Union

EN
L series

2024/1689

12.7.2024

REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 13 June 2024

laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)

Why 'Trustworthiness'?

Official Journal
of the European Union

EN
L series

2024/1689

12.7.2024

REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 13 June 2024

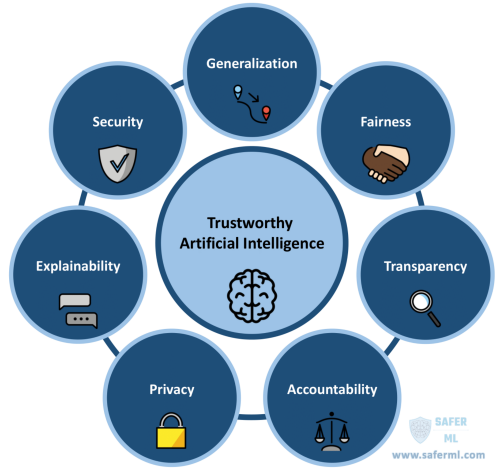
laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)

(1) The purpose of this Regulation is to improve the functioning of the internal market by laying down a uniform legal framework in particular for the development, the placing on the market, the putting into service and the use of artificial intelligence systems (AI systems) in the Union, in accordance with Union values, to promote the uptake of human centric and trustworthy artificial intelligence (AI) while ensuring a high level of protection of health, safety,

Trustworthiness in the AI Act - 7 Core Principles

"(27) The seven principles include human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being and accountab

1. **Human Agency:** Oversight & control.
2. **Technical Robustness:** Safety & resilience.
3. **Privacy:** Data governance.
4. **Transparency:** Traceability & communication.
5. **Fairness:** Diversity & non-discrimination.
6. **Well-being:** Social & environmental impact.
7. **Accountability:** Auditability.



The Translation Difficulty (I)

Human Trust (Sociology)

"The willingness to be vulnerable based on the expectation that the other party will perform a particular action..."

- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of management review, 20(3), 709-734.

Psychological State

The Translation Difficulty (I)

Human Trust (Sociology)

"The willingness to be vulnerable based on the expectation that the other party will perform a particular action..."

- *Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of management review, 20(3), 709-734.*

Psychological State

System Trustworthiness

A function of specific technical and non-technical attributes: Accuracy, Reliability, Resilience, ... + Agency, well-being ....

Measurable System Property

The Translation Difficulty (I)

Human Trust (Sociology)

"The willingness to be vulnerable based on the expectation that the other party will perform a particular action..."

- *Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of management review, 20(3), 709-734.*

Psychological State

System Trustworthiness

A function of specific technical and non-technical attributes: Accuracy, Reliability, Resilience, ... + Agency, well-being ....

Measurable System Property

The Semantic difficulty

We use a human-centric word to describe machine-centric behaviors. Operationalization is difficult because we are translating *intent* (vulnerability) into *benchmarks* (math).

The Operationalization difficulty (II)

The Challenge: How do we map qualitative values to quantitative systems?

The Operationalization difficulty (II)

The Challenge: How do we map qualitative values to quantitative systems?

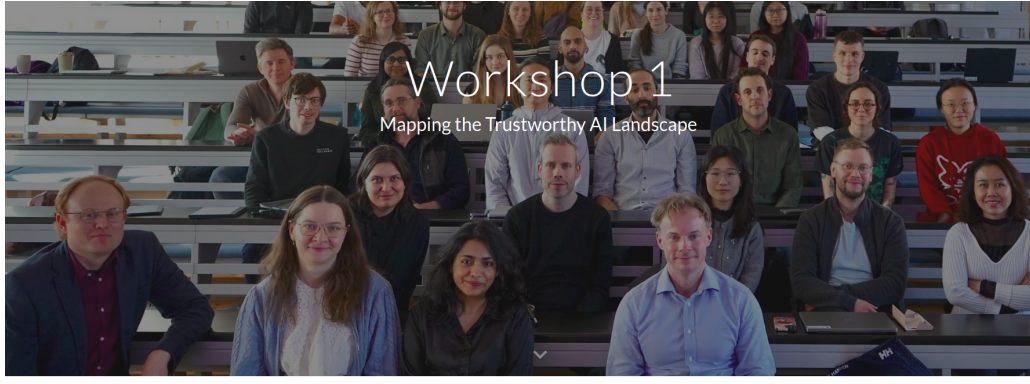
- **The Regulator's Language:** 'Ensure the system is fair and provides human agency.'

The Operationalization difficulty (II)

The Challenge: How do we map qualitative values to quantitative systems?

- **The Regulator's Language:** 'Ensure the system is fair and provides human agency.'
- **The Engineer's Language:** 'What is the loss function? Which statistical parity metric should I optimize for to ensure the system is fair?'

The Fractured Definition (III)



When: 3rd March 2026 11:30 - 16:00

Where: Pioneer Center for AI, Øster Voldgade 3, 1350 Copenhagen

Focus: Establish baseline understanding of trustworthy AI research across Denmark and Nordic countries, leading to a living repository of Trustworthy AI expertise in Denmark.

Figure: https://www.aicentre.dk/p1-programs/trustworthy-artificial-intelligence-trai
https://sites.google.com/view/trai-p1-program/home?authuser=0

The fractured Definitions (III)

What to map?

Ethics and Philosophy

Focuses on **moral agency** and social justice.
Asks if the system *should* be trusted.

The fractured Definitions (III)

What to map?

Ethics and Philosophy

Focuses on **moral agency** and social justice. Asks if the system *should* be trusted.

Software Engineering

Focuses on **reliability**. Trust as a byproduct of predictability.

The fractured Definitions (III)

What to map?

Ethics and Philosophy

Focuses on **moral agency** and social justice. Asks if the system *should* be trusted.

Computer Science

Focuses on **bounded risk** and mathematical guarantees for robustness, uncertainty, explainability, privacy ....

Software Engineering

Focuses on **reliability**. Trust as a byproduct of predictability.

The fractured Definitions (III)

What to map?

Ethics and Philosophy

Focuses on **moral agency** and social justice. Asks if the system *should* be trusted.

Software Engineering

Focuses on **reliability**. Trust as a byproduct of predictability.

Computer Science

Focuses on **bounded risk** and mathematical guarantees for robustness, uncertainty, explainability, privacy ....

Human Computer Interaction

Focuses on **subjective perception** and the human belief that a system will perform.

The fractured Definitions (III)

What to map?

Ethics and Philosophy

Focuses on **moral agency** and social justice. Asks if the system *should* be trusted.

Computer Science

Focuses on **bounded risk** and mathematical guarantees for robustness, uncertainty, explainability, privacy ....

Software Engineering

Focuses on **reliability**. Trust as a byproduct of predictability.

Human Computer Interaction

Focuses on **subjective perception** and the human belief that a system will perform.

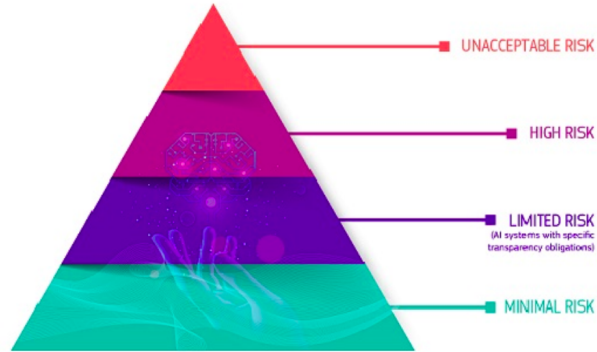
The silos

We are trying to build more **shared language and understanding** between these domains.

The Context of Risk (IV)

The EU AI Act uses a **risk-based approach**:

- **Unacceptable:** Prohibited (Social scoring).
- **High Risk:** Mandatory requirements (Hiring, Health, Justice).
- **Minimal:** Transparency only.



Application Drift → Context

Contextual Dependency

A model designed for one application could become untrustworthy when applied to another, even if the underlying mathematics and weights remain identical.

Example: LLM Chatbots

- **Application A (Creative Writing):**
  Low Stakes → Hallucinations are perceived as *'Creative Features.'*
- **Application B (Medical Advice):**
  High Stakes → Hallucinations are perceived as *'High impact Failures.'*

Changes in context:

When the **Application (Risk)** changes, the evaluation and auditing practices (metrics) could also change.

Moving Forward

1. **Unifying Domains:** Bridging the fractured definitions by creating a shared vocabulary between Ethics, Computer Science, and Engineering.

Moving Forward

1. **Unifying Domains:** Bridging the fractured definitions by creating a shared vocabulary between Ethics, Computer Science, and Engineering.
2. **Lifecycle Governance:** Adopting a unified view of the AI value chain $\rightarrow$ from data curation to real-world operational monitoring.

Moving Forward

1. **Unifying Domains:** Bridging the fractured definitions by creating a shared vocabulary between Ethics, Computer Science, and Engineering.
2. **Lifecycle Governance:** Adopting a unified view of the AI value chain → from data curation to real-world operational monitoring.
3. **Verifiable Evidence:** Invest in rigorous frameworks that allow us to measure, audit, and constrain system behavior.

The Path Ahead

Trustworthiness is not a final destination or a static metric; it is an evolving process of operational verification across the entire system lifecycle.