Towards Trustworthy AI

Integrating Reasonig and Learning

Fredrik Heintz

Dept. of Computer Science, Linköping University

fredrik.heintz@liu.se

@FredrikHeintz

LINKÖPINGS UNIVERSITET

ELLIIT

TrustLLM

Ethics Guidelines for Trustworthy AI

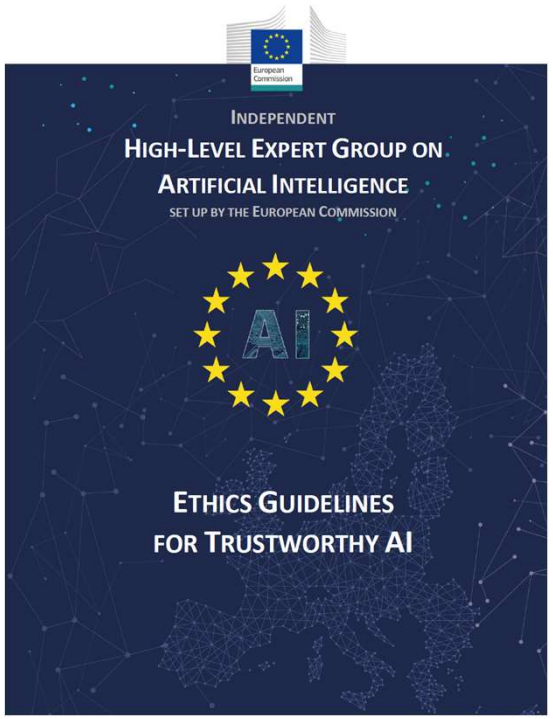
Human-centric approach: AI as a means, not an end

Trustworthy AI as our foundational ambition, with three components

- Lawful AI
- Ethical AI
- Robust AI

Three levels of abstraction

- from principles (Chapter I)
- to requirements (Chapter II)
- to assessment list (Chapter III)



https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Ethics Guidelines for Trustworthy AI

4 Ethical Principles based on fundamental rights

Respect for human autonomy

Augment, complement and empower humans

Prevention of harm

Safe and secure. Protect physical and mental integrity.

Fairness

Equal and just distribution of benefits and costs.

Explicability

Transparent, open with capabilities and purposes, explanations

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Ethics Guidelines for Trustworthy AI

Human agency and oversight

Technical Robustness and safety

Privacy and data governance

Transparency

Diversity, non-discrimination and fairness

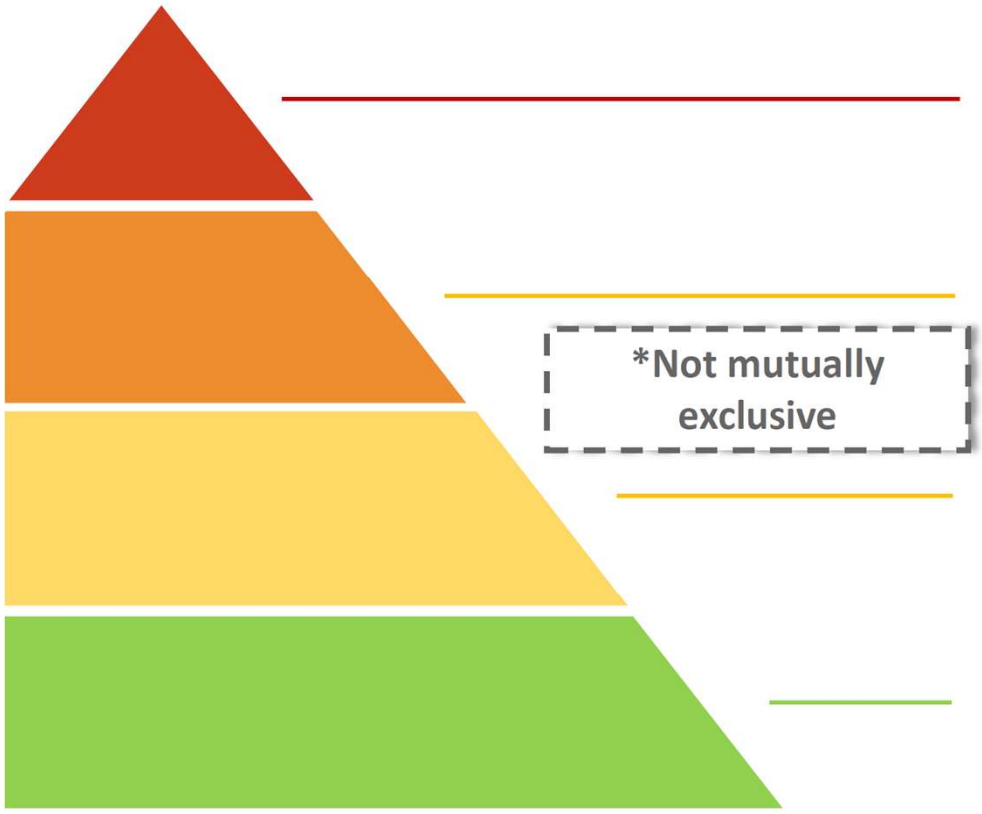
Societal & environmental well-being

Accountability

To be continuously implemented & evaluated throughout AI system's life cycle

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

A risk-based approach



*Not mutually exclusive

Unacceptable risk
e.g. social scoring
— **Prohibited**

High risk
e.g. recruitment, medical devices
— **Permitted** subject to compliance with AI requirements and ex-ante conformity assessment

'Transparency' risk
'Impersonation' (bots)
— **Permitted** but subject to information/transparency obligations

Minimal or no risk
— **Permitted** with no restrictions

TAILOR

Foundation of Trustworthy AI:
Integrating Learning, Optimisation and Reasoning

Fredrik Heintz

Dept. of Computer Science, Linköping University

fredrik.heintz@liu.se, @FredrikHeintz

https://tailor-network.eu/

TAILOR is an ICT-48 Network of AI Research Excellence Centers funded by EU Horizon 2020 research and innovation programme GA No 952215

LINKÖPINGS UNIVERSITET

TAILOR – Vision

Develop the scientific foundations for Trustworthy AI

integrating learning, optimisation and reasoning

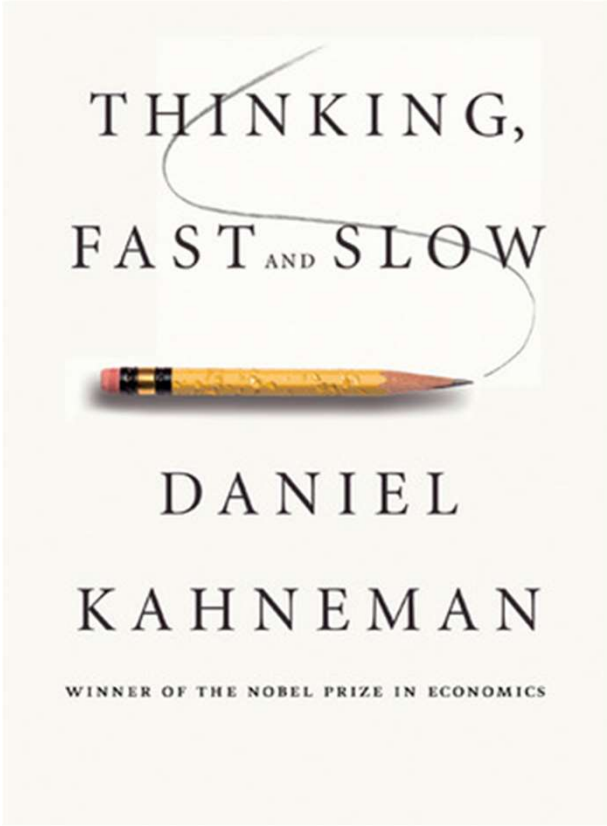
to realise the European vision of human-centered Trustworthy AI.

This project is funded by the EC under H2020 ICT-48

Human and Computational Thinking

Figure 1: A Comparison of System 1 and System 2 Thinking

System 1 "Fast"	System 2 "Slow"
DEFINING CHARACTERISTICS Unconscious Effortless Automatic	DEFINING CHARACTERISTICS Deliberate and conscious Effortful Controlled mental process
WITHOUT self-awareness or control	WITH self-awareness or control
"What you see is all there is."	Logical and skeptical
ROLE Assesses the situation Delivers updates	ROLE Seeks new/missing information Makes decisions



This project is funded by the EC under H2020 ICT-48

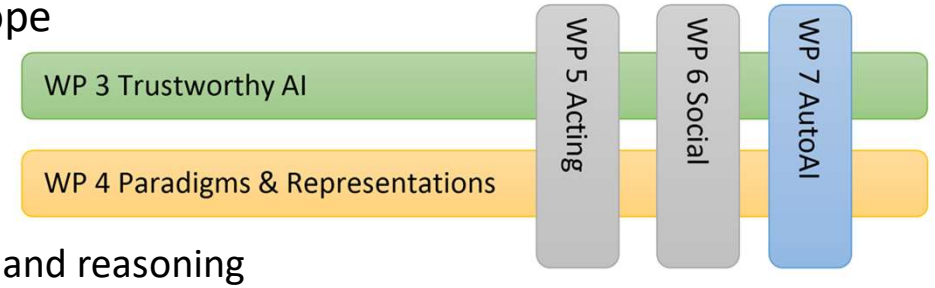
TAILOR ICT-48 Network

CLAIRE

TAILOR
https://tailor-network.eu/

*TAILOR brings together 54 leading AI research centres from **learning**, **optimisation** and **reasoning** together with major European companies representing important industry sectors into a single scientific network addressing the **scientific foundations of Trustworthy AI** to reduce the fragmentation, boost the collaboration, and increase the AI research capacity of Europe as well as attracting and retaining talents in Europe.*

- 54 research excellence centres from 20 countries across Europe coordinated by Fredrik Heintz, Linköping University, Sweden
- Four instruments
  - An ambitious research and innovation roadmap
  - Five basic research programs integrating learning, optimisation and reasoning in key areas for providing the scientific foundations for Trustworthy AI
  - A connectivity fund for active dissemination to the larger AI community
  - Network collaboration promoting research exchanges, training materials and events, and joint PhD supervision

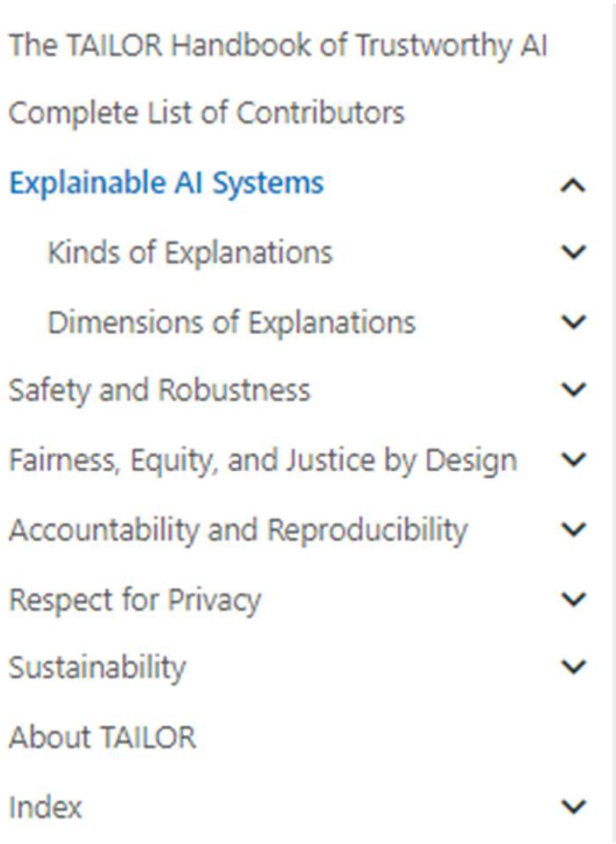


This project is funded by the EC under H2020 ICT-48

TAILOR
https://tailor-network.eu/

Trustworthy AI Handbook

- An **online encyclopedia** of the major scientific and technical terms related to Trustworthy AI
- Contains an overview of the **main dimensions of trustworthiness**, major challenges and solutions in the field, and the latest research developments
- For **non experts, researchers and students**
- 30 contributors from all areas of Trustworthy AI
- Integrated process for enrichment of Wikipedia while maintaining the integrity of the Handbook
- 1st version available: https://tailor-network.eu/handbook/



This project is funded by the EC under H2020 ICT-48

TAILOR
https://tailor-network.eu/

STRATEGIC RESEARCH & INNOVATION ROADMAP OF TRUSTWORTHY AI

The Scientific Foundations of Trustworthy AI in Europe for the Years 2022-2030

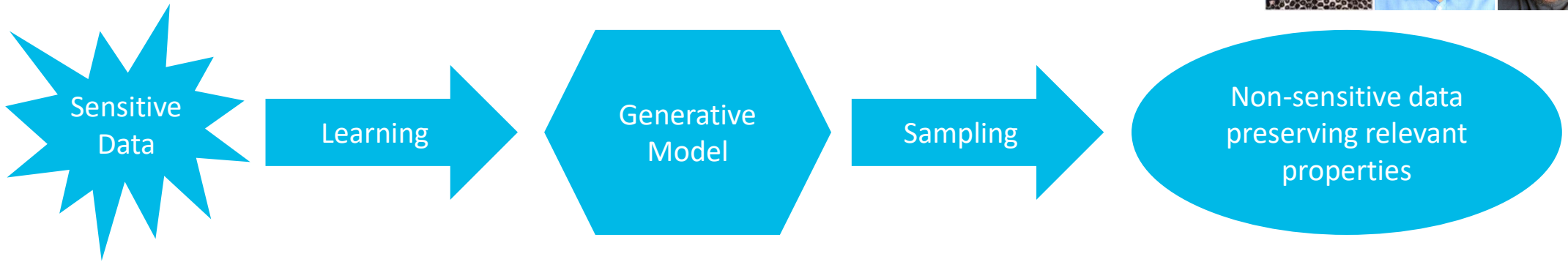
TAILOR Strategic Research and Innovation Roadmap (SRIR) aims to boost research on Trustworthy AI by clearly defining the major research challenges.

https://tailor-network.eu/research-overview/strategic-research-and-innovation-roadmap/

This project is funded by the EC under H2020 ICT-48

Privacy-preserving synthetic data generation

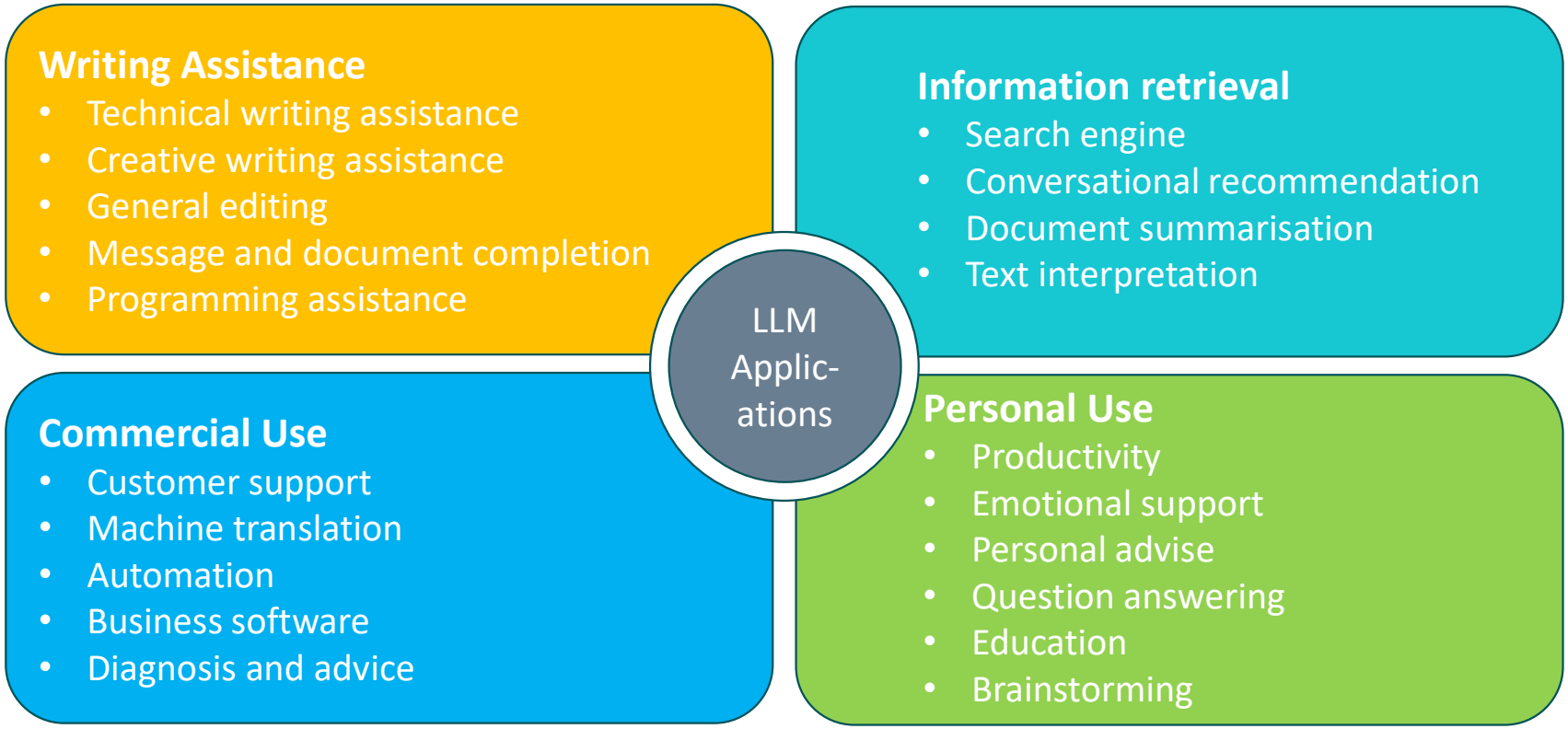
[R. Ramachandranpillai, Md F. Sikder, D. Bergström]



1. Learn a generative model that captures the probability distribution of the sensitive data
2. Create a synthetic data set from the generative model that both captures the salient features of the original data set **and** is non-sensitive
3. Methods for verifying that the synthetic data set is accurate enough
4. Methods for verifying that the synthetic data set is non-sensitive

Fair Latent Deep Generative Models (FLDGMs) for Syntax-Agnostic and Fair Synthetic Data Generation, *Resmi Ramachandranpillai*, Md Fahim Sikder*, Fredrik Heintz*, ECAI23
Bt-GAN: Generating Fair Synthetic Healthdata via Bias-transforming Generative Adversarial Networks, *Resmi Ramachandranpillai, Md Fahim Sikder, David Bergström, Fredrik Heintz,* Accepted to JAIR.

Large Language Model Applications



Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment by Yang Liu Etal, 2023

Can you Trust ChatGPT? No!

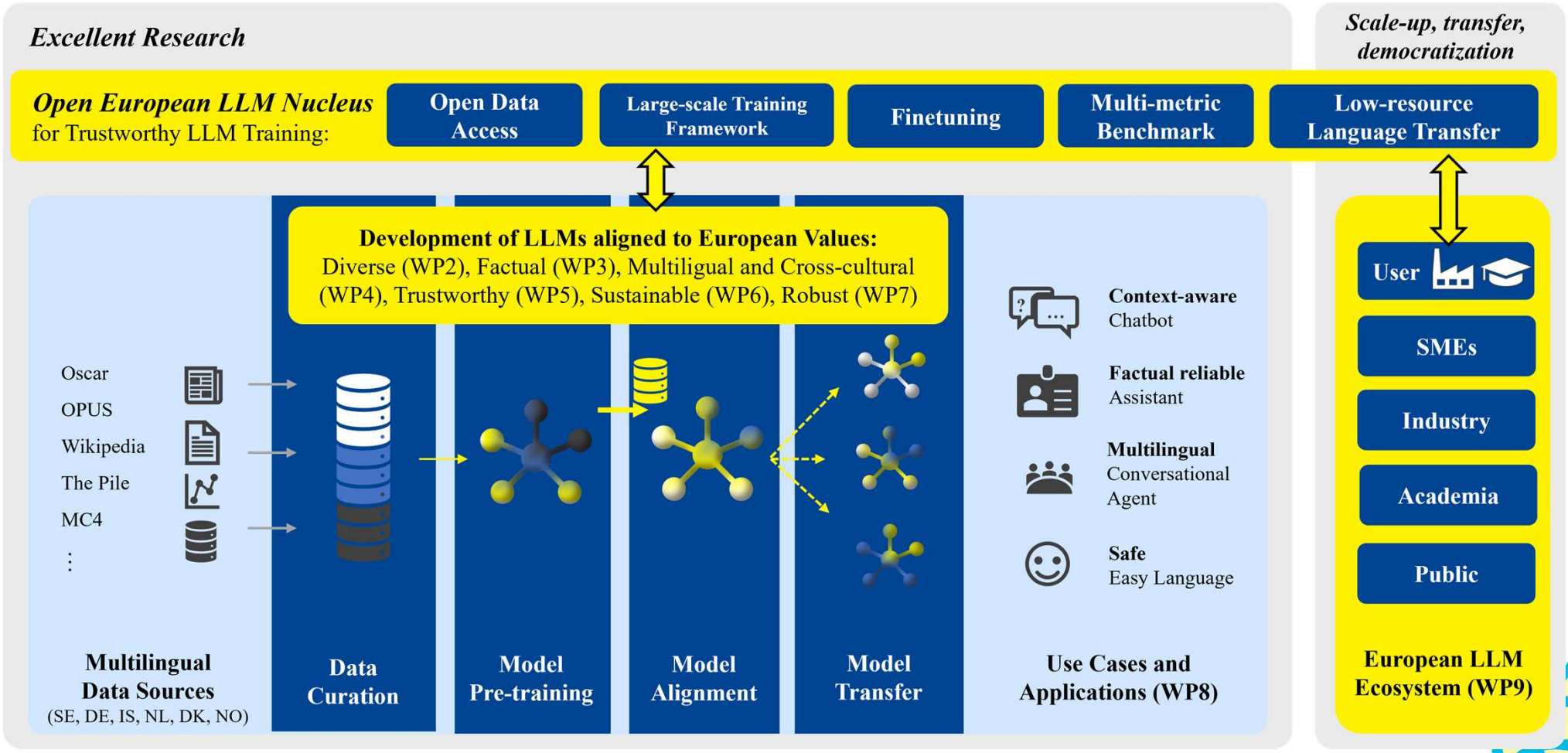
- Very limited information about the training data
- It makes things up, with confidence (hallucinations)
- Even when there are references these may be false or not applicable
- Cannot count or draw logical conclusions
- Stuck in time
- *but, ChatGPT is still useful!*

TrustLLM – Trustworthy and Factual LLMs made in Europe

- Develop an open, trustworthy, and sustainable LLM initially targeting the Germanic languages.
- TrustLLM will tackle the full range of challenges of LLM development,
  - from ensuring sufficient quality and quantity of multilingual training data,
  - to sustainable efficiency and effectiveness of model training,
  - to enhancements and refinements for factual correctness, transparency, and trustworthiness,
  - to a suite of holistic evaluation benchmarks validating the multi-dimensional objectives.

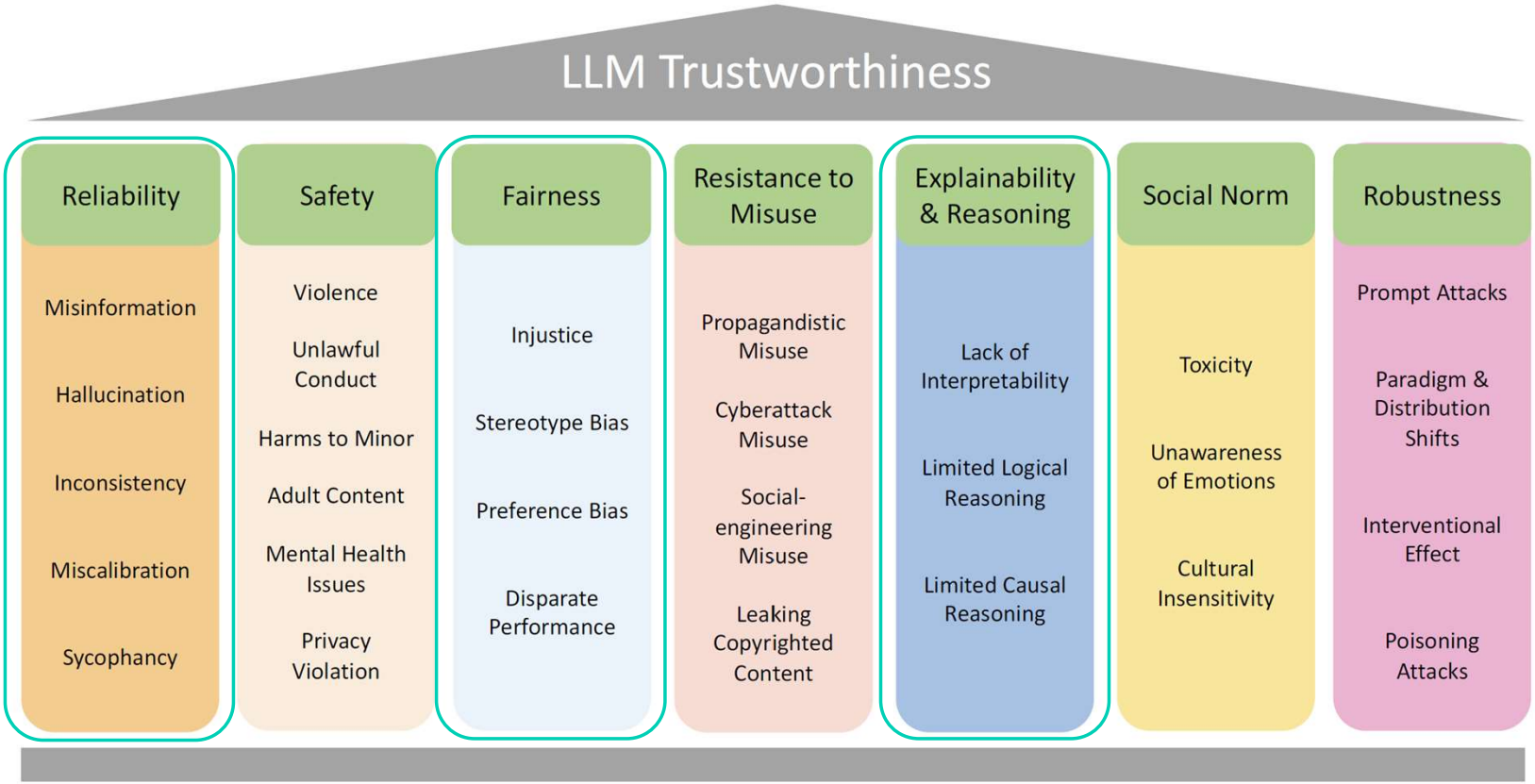
LINKÖPING UNIVERSITY
MIÐEIND
ALEXANDRA INSTITUTTET
Fraunhofer
UNIVERSITY OF ICELAND
UNIVERSITY OF COPENHAGEN
JÜLICH Forschungszentrum
TNO
KI BUNDESVERBAND
Lindholmen Science Park
AI SWEDEN
NTNU | Norwegian University of Science and Technology
Funded by the European Union

TrustLLM: Project Concept



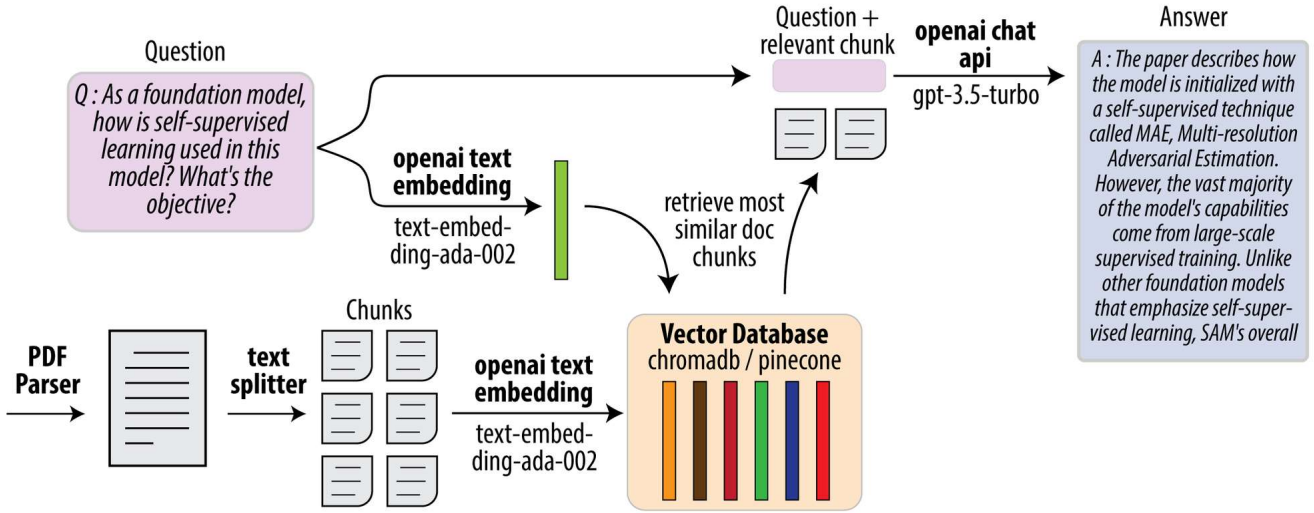
Funded by the European Union

LLM Trustworthiness



TrustLLM Improving Factual Correctness

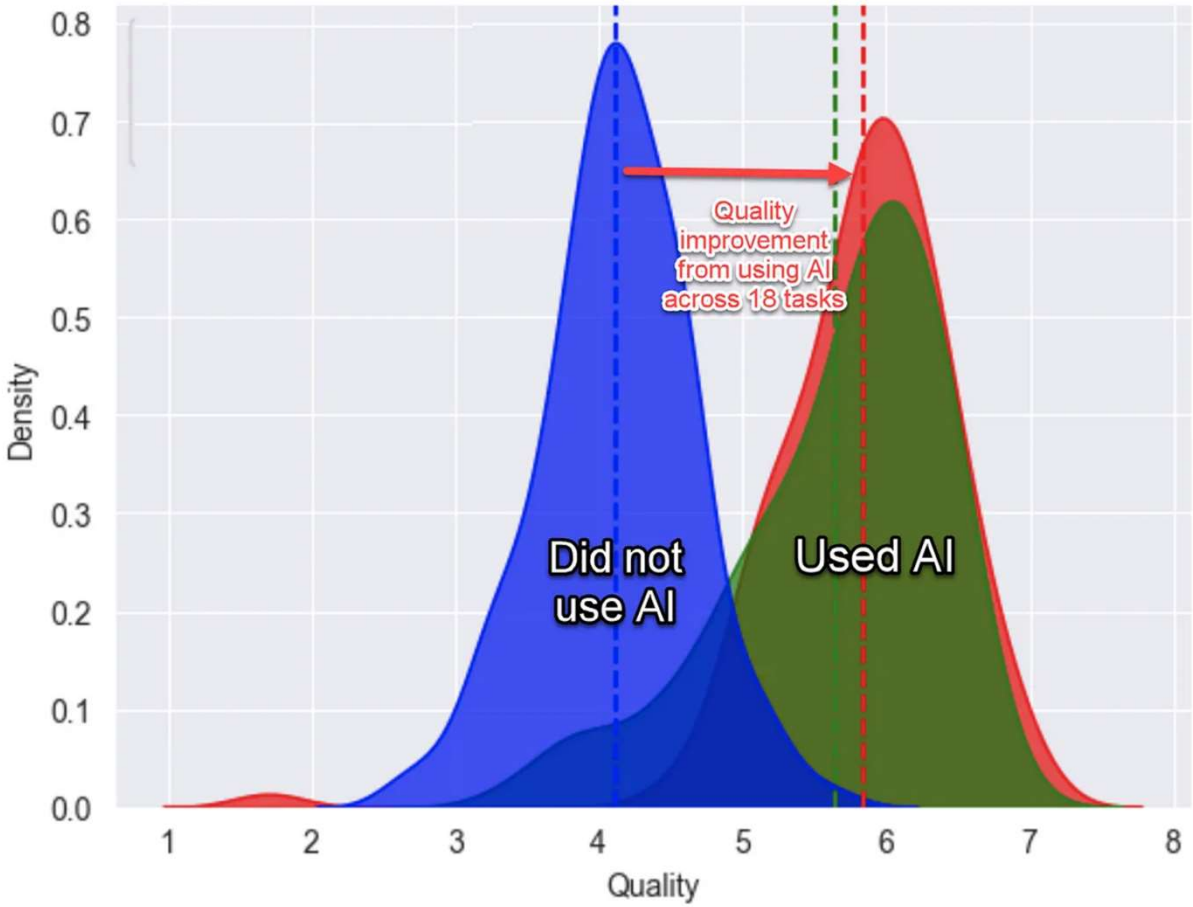
- Improving the **factual correctness** of LLMs concerning **static knowledge** (e.g. facts such as "*The Transformer architecture was invented by Google*"),
- Improving the **factual correctness** concerning **dynamic knowledge** (e.g., "*Ronaldo played for Real Madrid in 2017 and for Juventus Turin in 2018"*), and
- Improving the **multi-step** (common sense) **reasoning** capabilities of LLMs to **reason across information sources**.



Funded by the European Union

AI and Future of Work

- **12% more tasks finished**
- **25% quicker completion**
- **40% higher quality**



Distribution of output quality across all the tasks. The blue group did not use AI, the green and red groups used AI, the red group got some additional training on how to use AI.

https://www.oneusefulthing.org/p/centaurs-and-cyborgs-on-the-jagged

HBS Working Paper 24-013 "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality" by F. Dell'Acqua et al.

Take Away Message

- AI is about understanding intelligence and develop systems that exhibit intelligent behavior.
- AI will affect all aspects of our society. **Trust is essential!**
- To be **trustworthy** an **AI-system** should be **legal**, **ethical** and **robust**.
- Europe has **many initiatives** in the area, but **more** is needed.
- Several important research challenges remain such as
  - safety/robustness,
  - explainability/interpretability,
  - fairness/equity/justice, and
  - governance/accountability
- Very active and interdisciplinary research problems that are still mostly unsolved.
- **The TAILOR project is committed to develop the scientific foundations for Trustworthy AI**
- **Will most likely require integrating model-free data-driven learning approaches with model-based knowledge-driven reasoning approaches**