ETHICS AND TRUST FROM A HUMAN-CENTRIC PERSPECTIVE

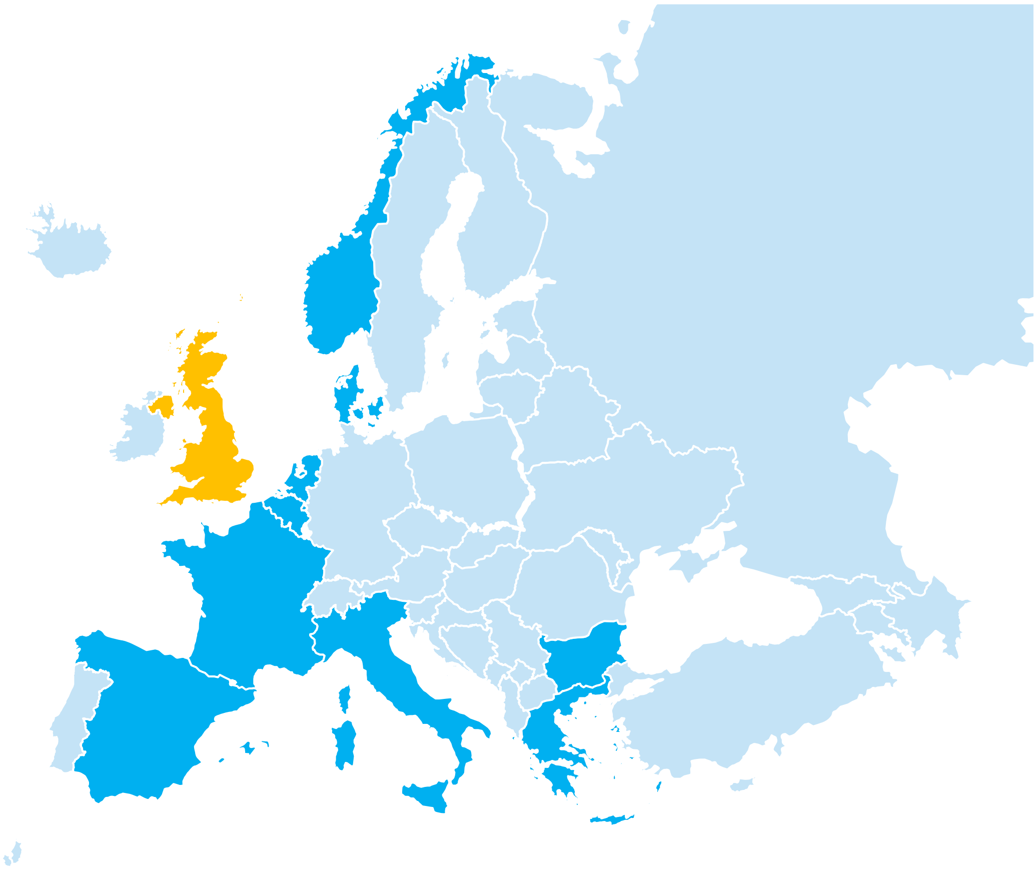
ENFIELD WEBINAR, NOVEMBER 15, 2024

ASBJØRN FØLSTAD, SINTEF

THEMIS 5.0 (RIA): Human-centred trustworthiness optimization in hybrid decision making support

Call: HORIZON-CL4-2022-HUMAN-02 - 'a human-centred and ethical development of digital and industrial technologies'.

Topic: HORIZON-CL4-2022-HUMAN-02-01 AI for human empowerment (AI, Data and Robotics Partnership)

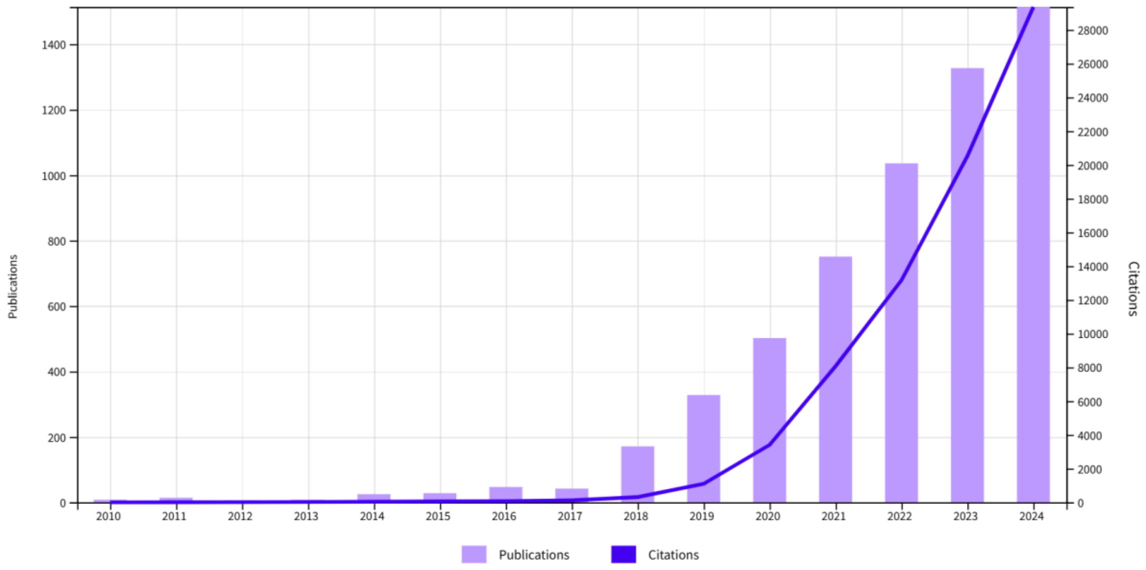


ETHICS, TRUST, AND TRUSTWORTHY AI

- **AI ethics and trust** key issues as AI becomes increasingly advanced and taken up in society

- **Trustworthy AI** key to meeting ethical requirements and establish a sound basis for trust in AI

Citation report: ethics AND «artificial intelligence»

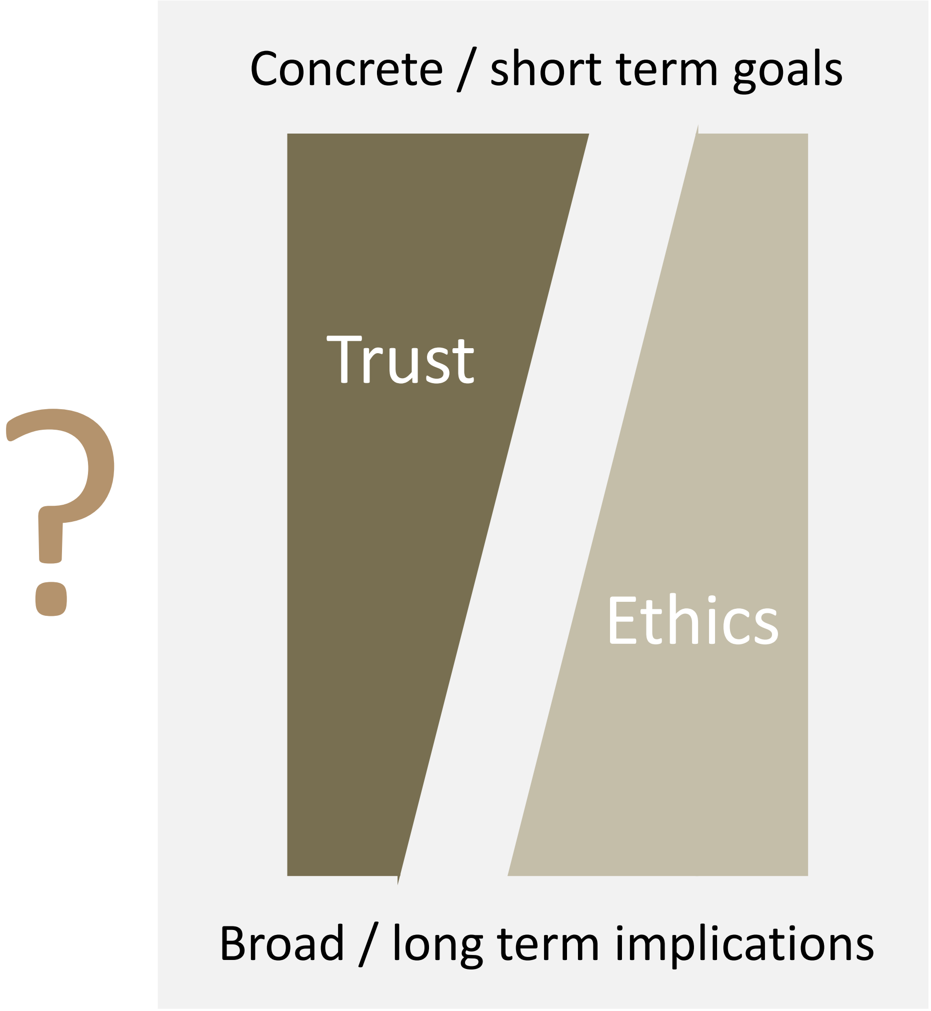


https://www.webofscience.com/

ETHICS, TRUST, AND TRUSTWORTHY AI

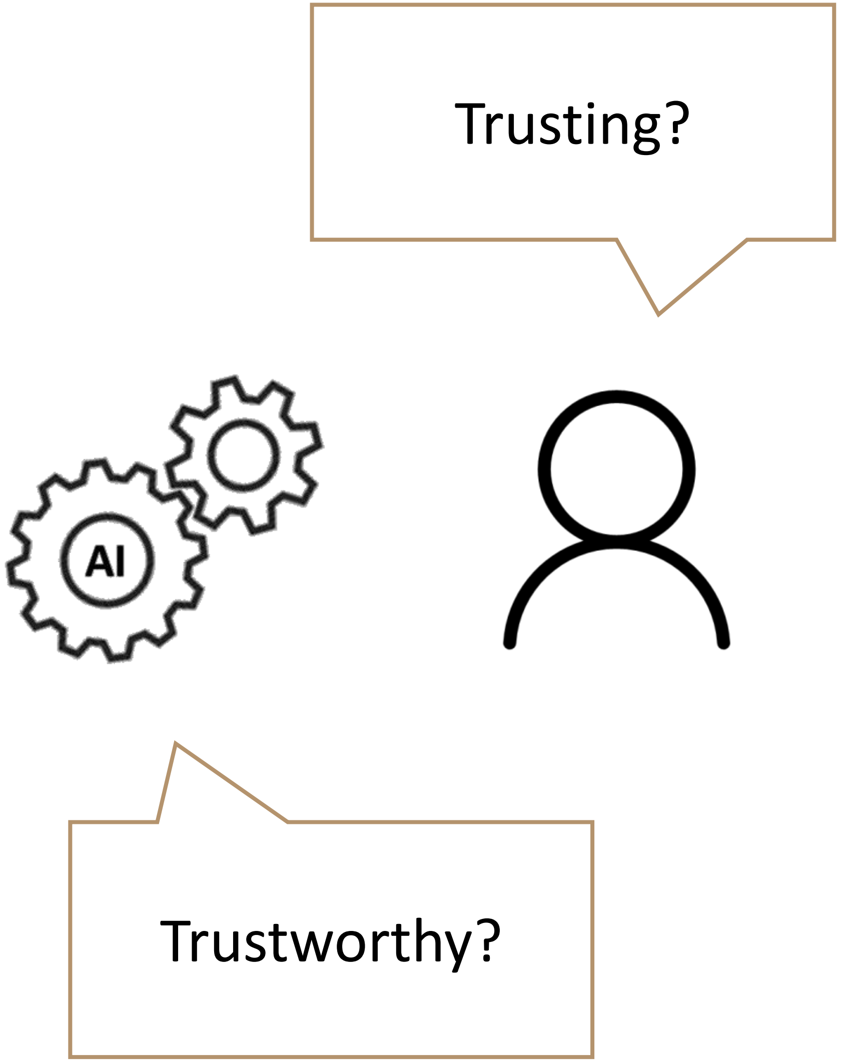
- **AI ethics and trust** key issues as AI becomes increasingly advanced and taken up in society

- **Trustworthy AI** key to meeting ethical requirements and establish a sound basis for trust in AI



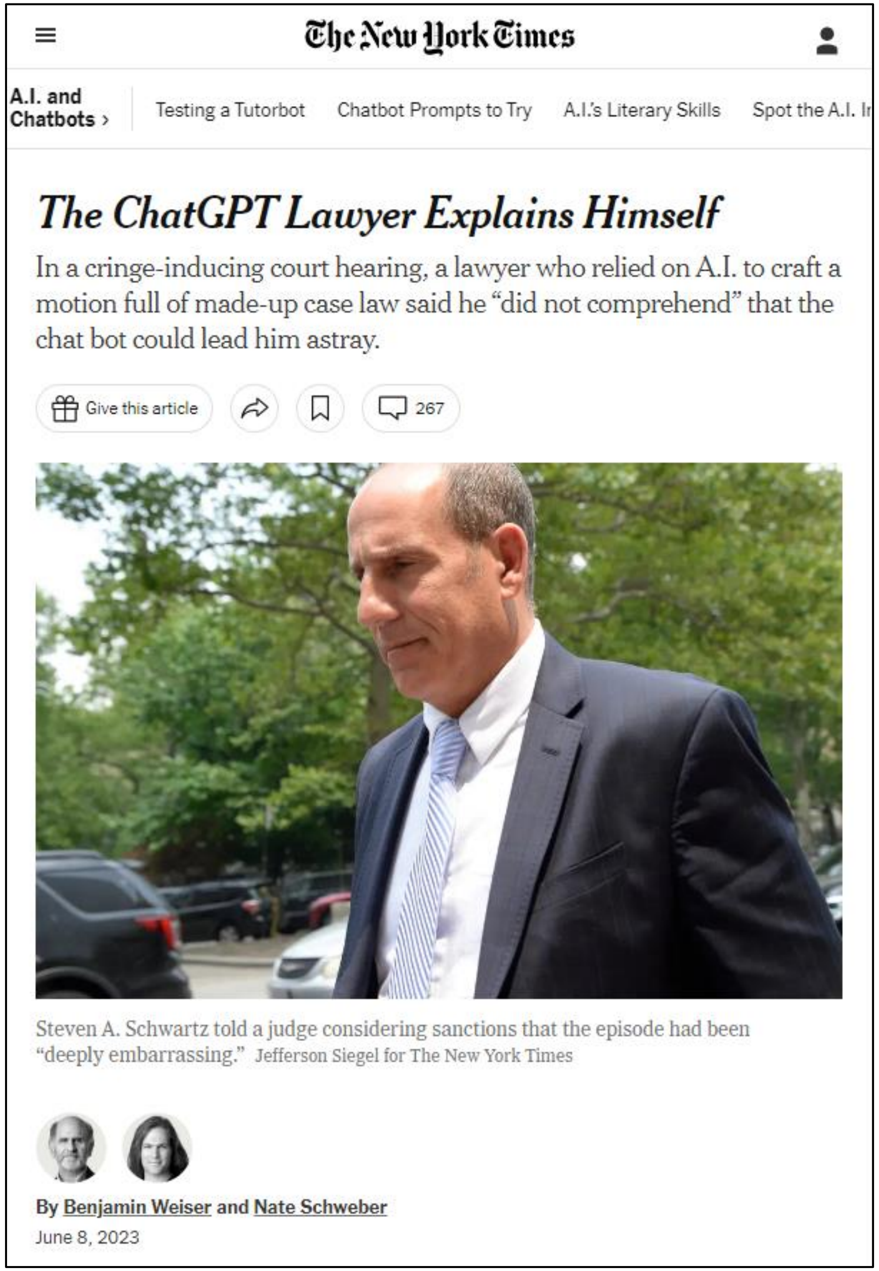
TRUST AND TRUSTWORTHINESS

- Trust – in the eye of the beholder
- Trustworthiness – characteristic of the AI system



TRUST AND TRUSTWORTHINESS

- Trust – in the eye of the beholder
- Trustworthiness – characteristic of the AI system
- Beneficial use of AI depends on **adequate configuration of trust** to the actual trustworthiness of the AI system



The ChatGPT Lawyer Explains Himself

In a cringe-inducing court hearing, a lawyer who relied on A.I. to craft a motion full of made-up case law said he "did not comprehend" that the chat bot could lead him astray.

Steven A. Schwartz told a judge considering sanctions that the episode had been "deeply embarrassing." Jefferson Siegel for The New York Times

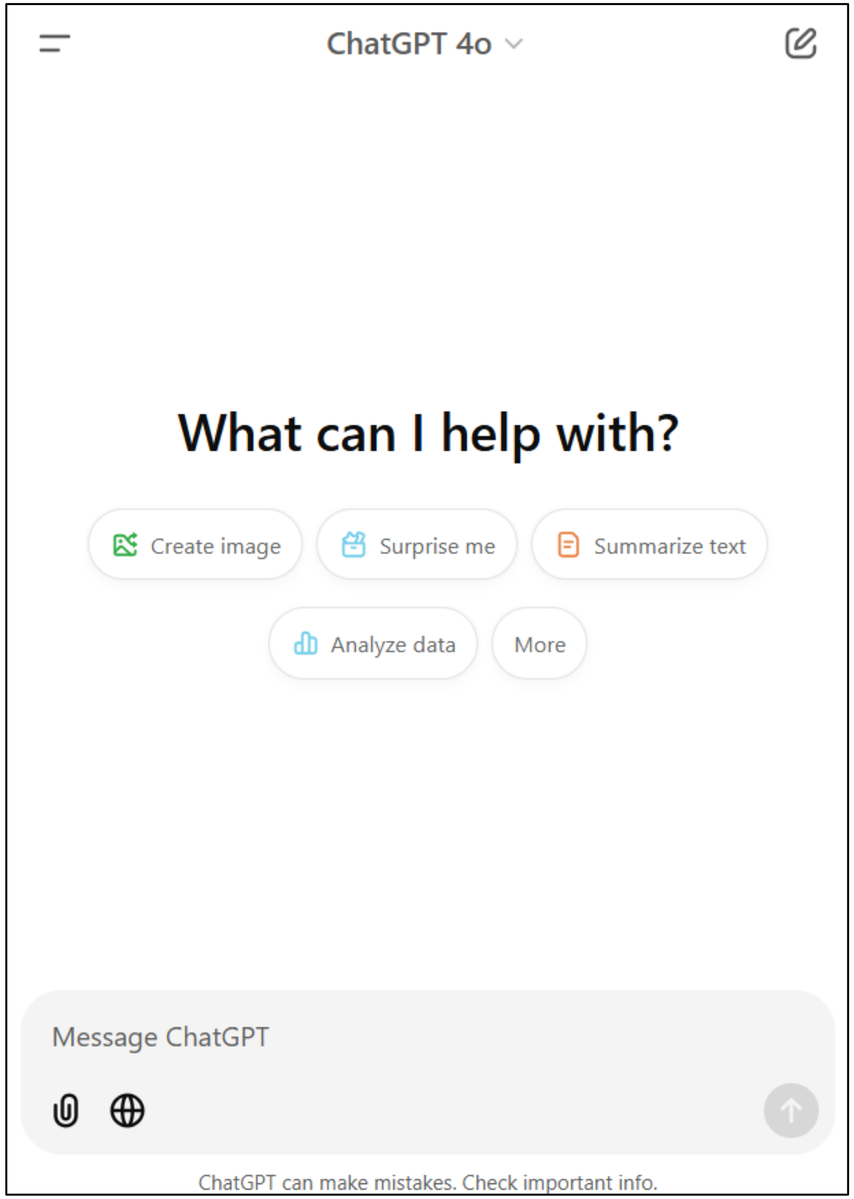
By Benjamin Weiser and Nate Schweber

June 8, 2023

hhttps://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html

TRUSTWORTHY AI

- Trustworthy AI addressed through **detectable trustworthiness characteristics**,
- … that is, characteristics on which adherence to ethical requirements can be assessed



ChatGPT 4o

What can I help with?

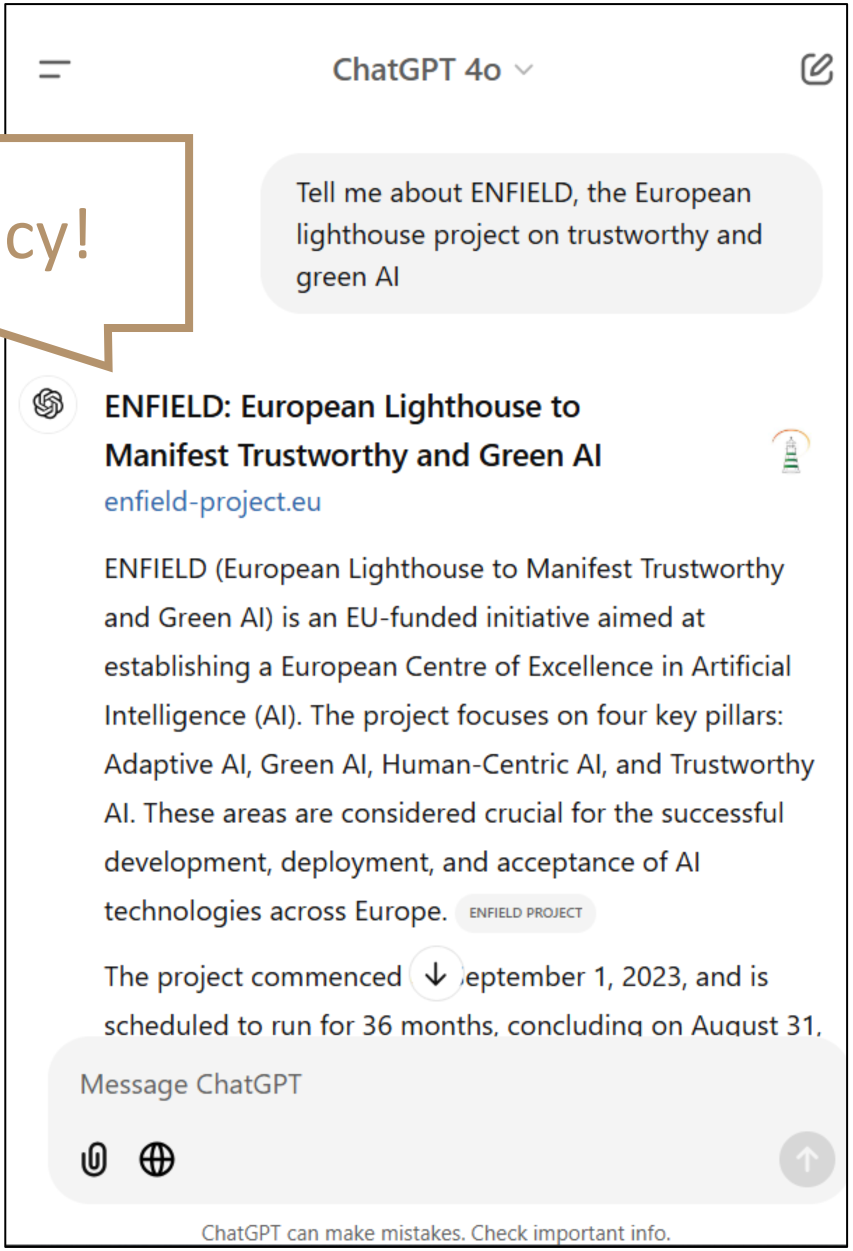
Message ChatGPT

ChatGPT can make mistakes. Check important info.

TRUSTWORTHY AI

- Trustworthy AI addressed through **detectable trustworthiness characteristics**,
- … that is, characteristics on which adherence to ethical requirements can be assessed

Accuracy!



Tell me about ENFIELD, the European lighthouse project on trustworthy and green AI

ENFIELD: European Lighthouse to Manifest Trustworthy and Green AI
enfield-project.eu

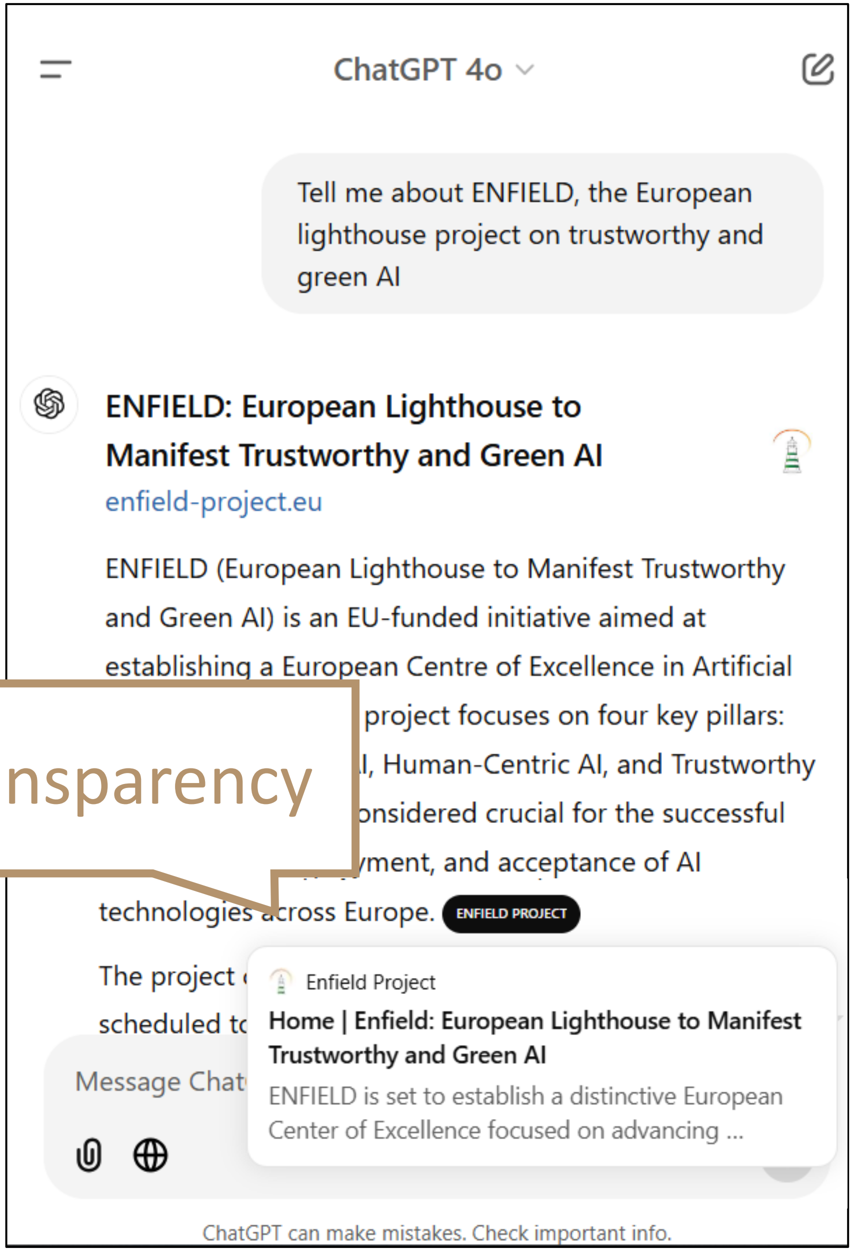
ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI) is an EU-funded initiative aimed at establishing a European Centre of Excellence in Artificial Intelligence (AI). The project focuses on four key pillars: Adaptive AI, Green AI, Human-Centric AI, and Trustworthy AI. These areas are considered crucial for the successful development, deployment, and acceptance of AI technologies across Europe. ENFIELD PROJECT

The project commenced ↓ eptember 1, 2023, and is scheduled to run for 36 months, concluding on August 31,

ChatGPT can make mistakes. Check important info.

TRUSTWORTHY AI

- Trustworthy AI addressed through **detectable trustworthiness characteristics**,
- … that is, characteristics on which adherence to ethical requirements can be assessed



Tell me about ENFIELD, the European lighthouse project on trustworthy and green AI

ENFIELD: European Lighthouse to Manifest Trustworthy and Green AI

enfield-project.eu

ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI) is an EU-funded initiative aimed at establishing a European Centre of Excellence in Artificial … project focuses on four key pillars: …l, Human-Centric AI, and Trustworthy …onsidered crucial for the successful …ment, and acceptance of AI technologies across Europe.

The project … scheduled to…

Enfield Project

Home | Enfield: European Lighthouse to Manifest Trustworthy and Green AI

ENFIELD is set to establish a distinctive European Center of Excellence focused on advancing …

ChatGPT can make mistakes. Check important info.

Transparency

ETHICS AND TRUSTWORTHINESS – CONCEPTUAL BALANCING

- **Balancing different approaches and frameworks**
- Ethics: Broad range of ethical guidelines and requirements
- Trustworthy AI: Broad range of trustworthiness characteristics. Partially overlapping frameworks

ETHICS AND TRUSTWORTHINESS – CONCEPTUAL BALANCING

- Balancing different approaches and frameworks
- **Ethics: Broad range of ethical guidelines and requirements**
- Trustworthy AI: Broad range of trustworthiness characteristics. Partially overlapping frameworks

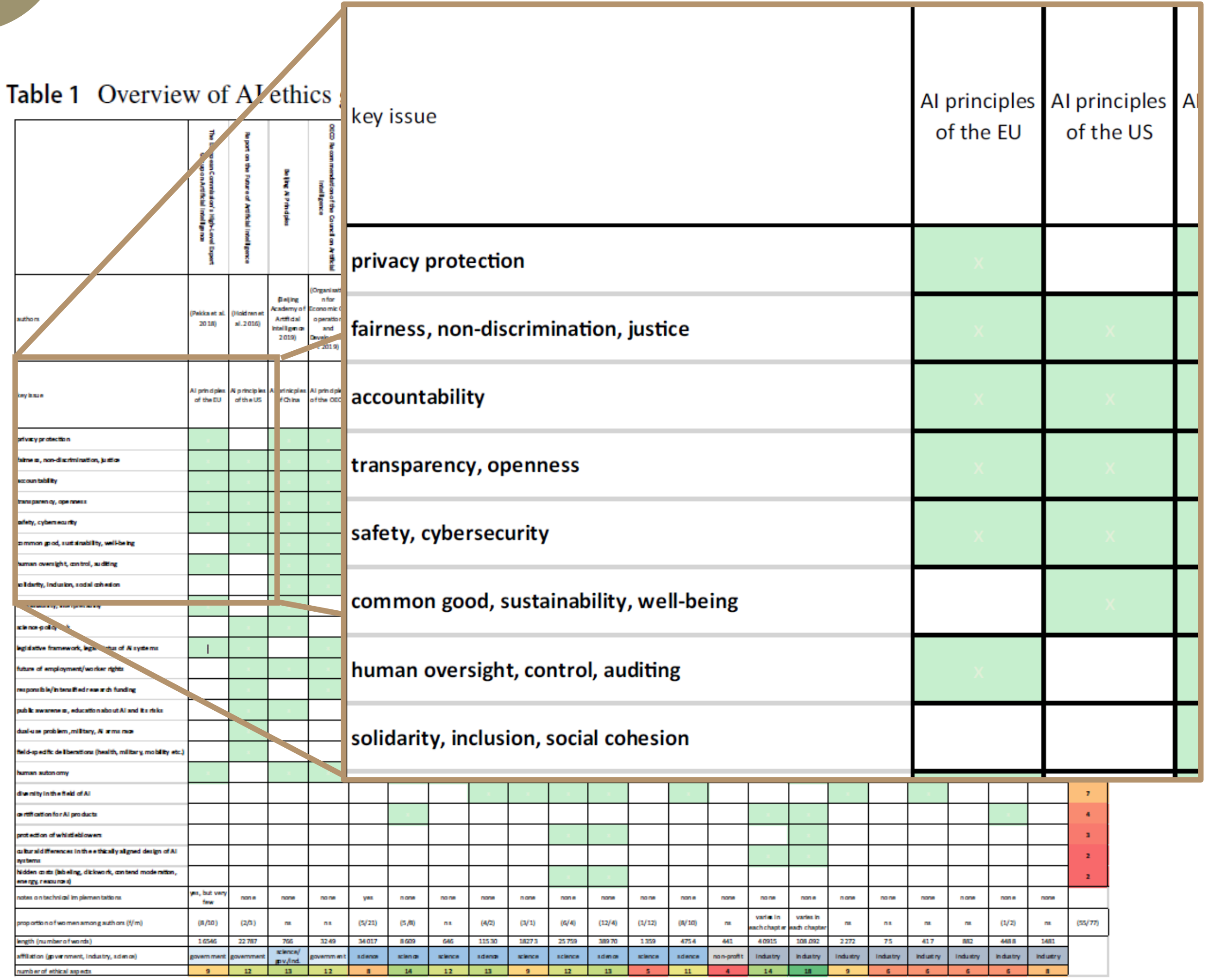


Table 1 Overview of AI ethics

key issue	AI principles of the EU	AI principles of the US
privacy protection	x	
fairness, non-discrimination, justice	x	x
accountability	x	x
transparency, openness	x	x
safety, cybersecurity	x	x
common good, sustainability, well-being		x
human oversight, control, auditing	x	
solidarity, inclusion, social cohesion		

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and machines, 30(1), 99-120.

ETHICS AND TRUSTWORTHINESS – CONCEPTUAL BALANCING

- Balancing different approaches and frameworks
- Ethics: Broad range of ethical guidelines and requirements
- **Trustworthy AI: Broad range of trustworthiness characteristics. Partially overlapping frameworks**

Human agency
Robustness and safety
Privacy
Transparency
Fairness
Well-being
Accountability

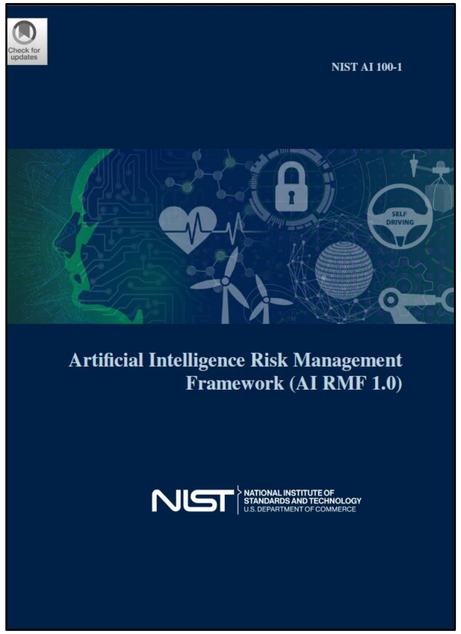
European Commission

INDEPENDENT
HIGH-LEVEL EXPERT GROUP ON
ARTIFICIAL INTELLIGENCE
SET UP BY THE EUROPEAN COMMISSION

AI

ETHICS GUIDELINES
FOR TRUSTWORTHY AI

Validity/reliability
Safety
Security/resilience
Explainability
Privacy
Fairness
Accountability


Check for updates

NIST AI 100-1

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

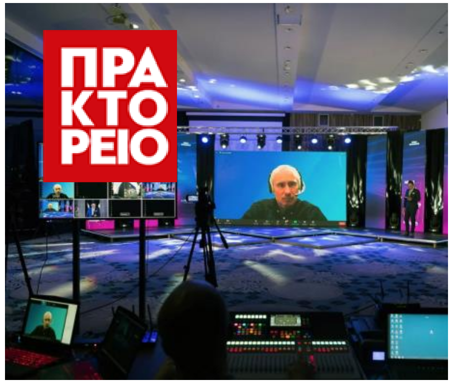
ETHICS AND TRUSTWORTHINESS – CONTEXTUAL BALANCING

- **Balancing for specific contexts and situations**
- Use case specific ethical requirements, with implications for AI trustworthiness assessment and optimization

ETHICS AND TRUSTWORTHINESS – CONTEXTUAL BALANCING

THEMIS 5.0 USE CASES:

- Balancing for specific contexts and situations
- **Use case specific ethical requirements, with implications for AI trustworthiness assessment and optimization**



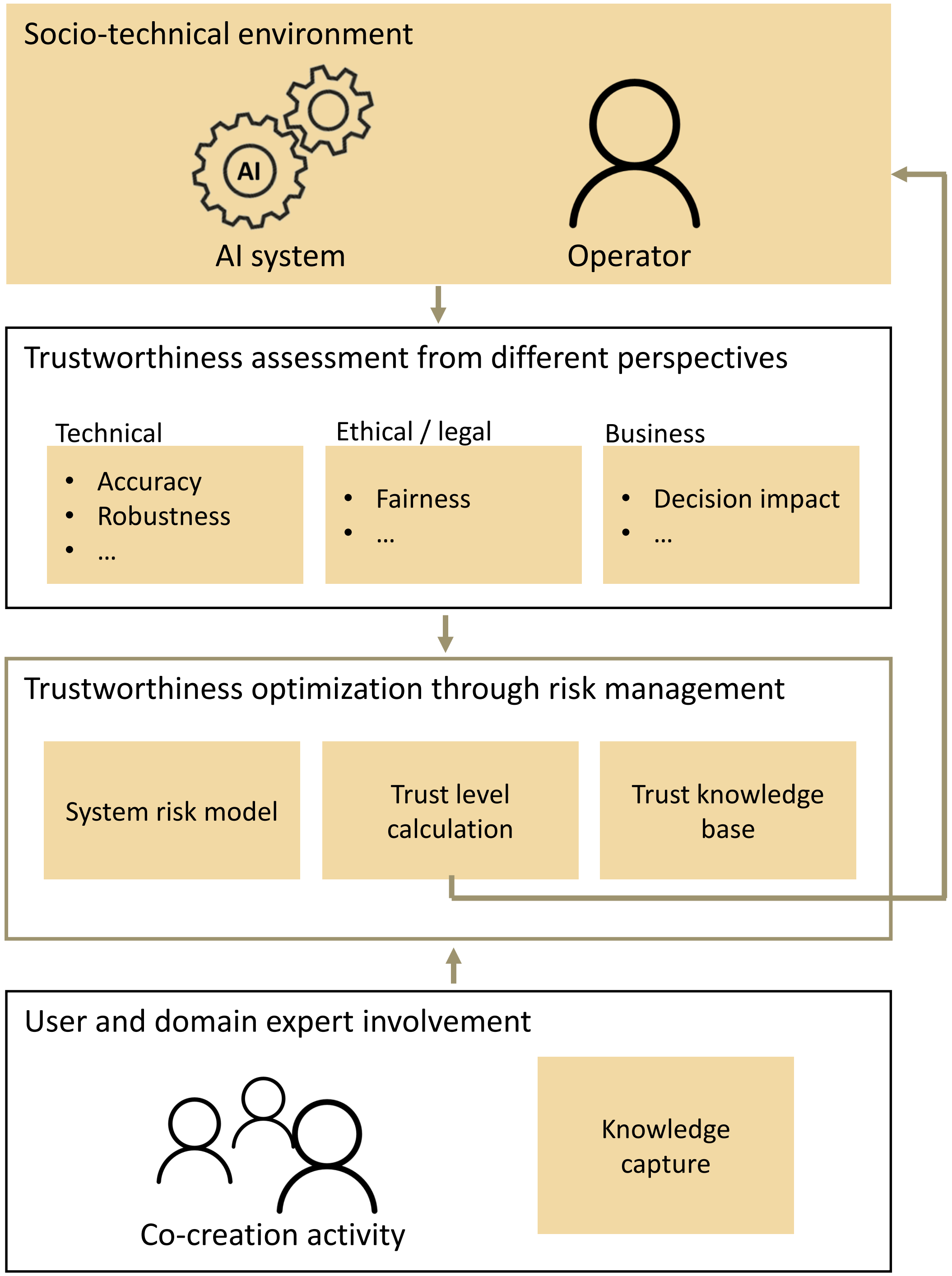
News media: Mitigate disinformation

Maritime: Optimise port operations

Healthcare: Personalised risk prediction

TRUSTWORTHINESS ASSESSMENT AND OPTIMIZATION – THE THEMIS 5.0 APPROACH

- Assessment of AI system in socio-technical environment
  - Use-cases: healthcare, news media, port management
- Detectable trustworthiness characteristics for assessment
- Trustworthiness optimization through risk management
  - Optimized with regards to user group preferences
- Human-centred assessment
  - Identification of trustworthiness characteristics with users and domain experts
  - Explanations for assessment by users and domain experts



Socio-technical environment

AI system

Operator

Trustworthiness assessment from different perspectives

Technical
- Accuracy
- Robustness
- …

Ethical / legal
- Fairness
- …

Business
- Decision impact
- …

Trustworthiness optimization through risk management

System risk model

Trust level calculation

Trust knowledge base

User and domain expert involvement

Co-creation activity

Knowledge capture

IN CONCLUSION

- AI ethics and trust with high attention, in society and academia alike
- Trustworthy AI key to meeting ethical requirements and establish a sound basis for trust in AI
- Ethical requirements and corresponding trustworthiness characteristics dependent on user case
- Human-centre approach needed to assess and optimize AI trustworthiness

THEMIS 5.0