Trustworthy AI, bias and the role of hybrid intelligence

Roberta Calegari

Alma Mater Studiorum–Università di Bologna, Italy

AI, Data, Robotics Forum #ADRF23
Versailles, Paris

09 November 2023

AEQUITAS

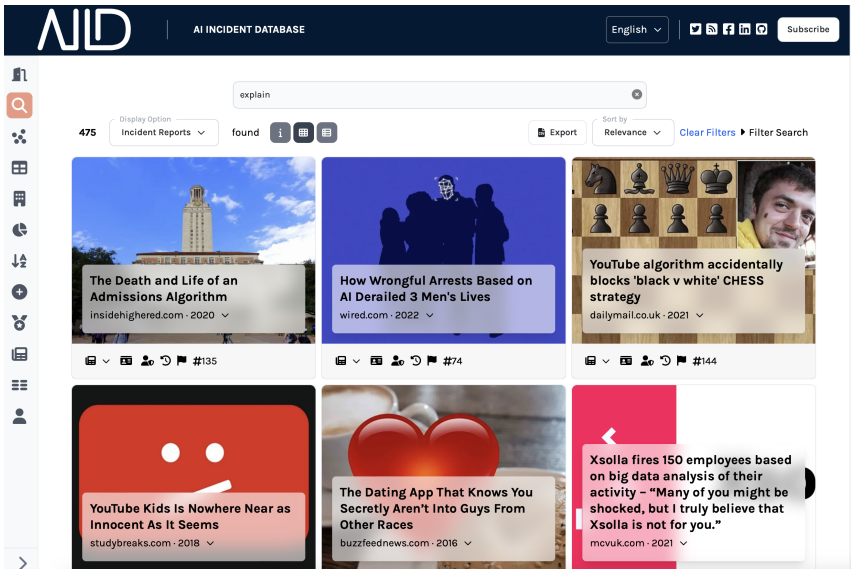
This work was supported by European Union's Horizon Europe research and innovation programme under grant number 101070363

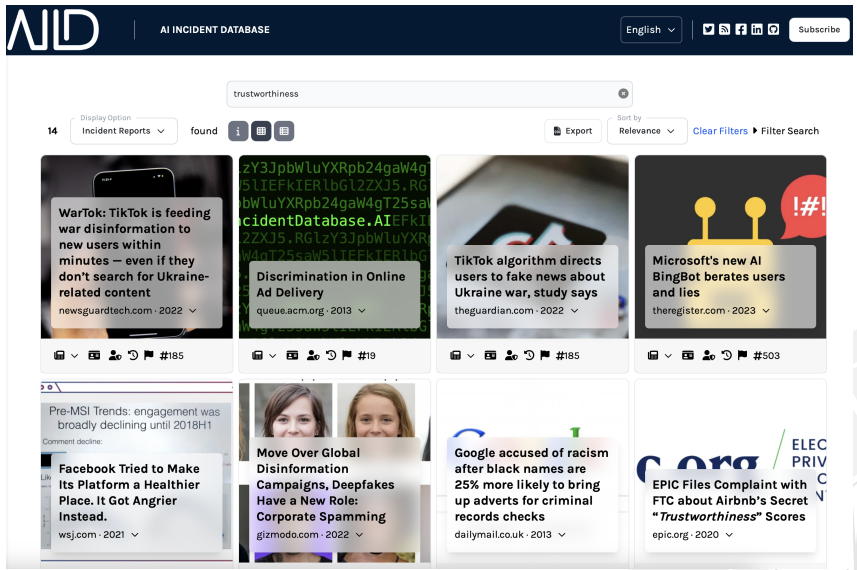
Context

AI systems to formalize, scale, and accelerate processes

- in cars to avoid accidents
- in banks to manage investments and loan decisions
- in hospitals to aid doctors in diagnosing and detecting disease
- in law enforcement to help officials recover evidence and make law enforcement easier
- in the military of many countries
- in insurance organizations to determine risk
- ...

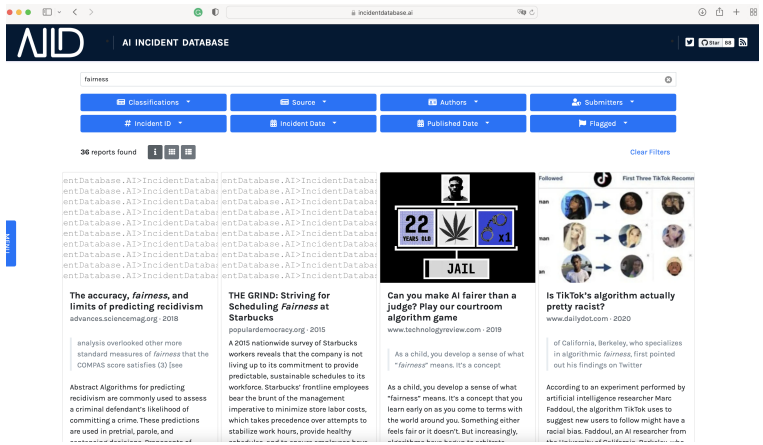
Needs: the reality





Why trustworthiness? Why bias?

- society is facing a dramatic increase in *pervasive inequality* and *intersectional discrimination* due to the widespread use of AI
  [Leavy et al., 2021, Leavy et al., 2020]
  - ML is contributing to creating a society where some groups or individuals are disadvantaged
- https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- https://www.technologyreview.com/s/610634/microsofts-neo-nazi-sexbot-was-a-great-lesson-for-makers-of-ai-assistants/



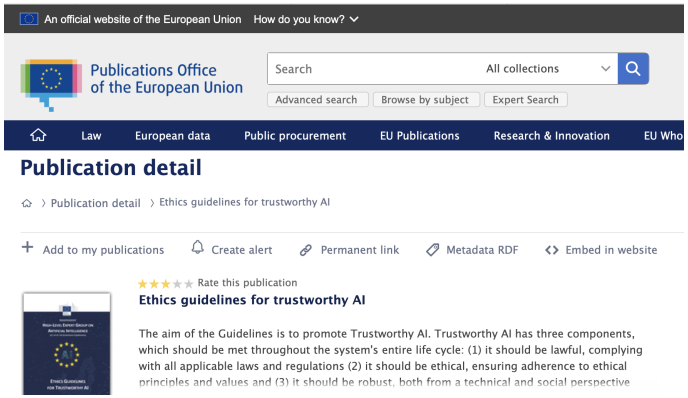
Why trustworthiness? Why bias?

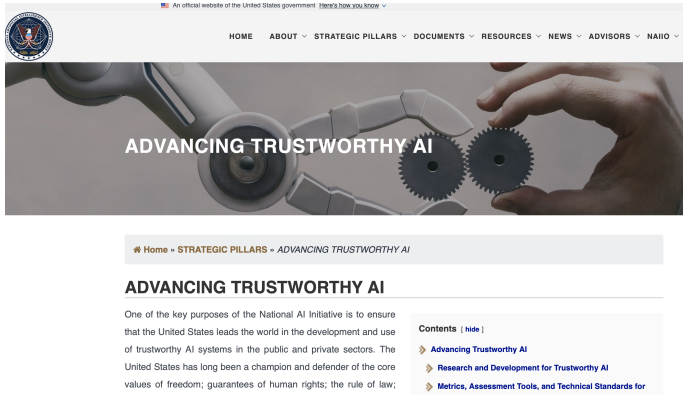
⇒ challenging to comprehend and trust AI outcomes

- black-box nature

⇒ understand the reasoning behind an AI model's decision-making

Trustworthy AI: initiatives





Europe Strategy

- Ethics Guidelines for Trustworthy AI (EG-TAI) [European Commission, 2019]
- First AI regulation (the "AI Act", 2021) [Act, 2021]
  - *ensuring* that AI systems, introduced on the *EU market* are trustworthy
  - creating *legal certainty* to facilitate investments and innovation in AI
- TAI is the basis for the development, deployment and use of AI in Europe

$\Rightarrow$ **close the AI "trust gap"**

EG-TAI: TAI Requirements & AI Act

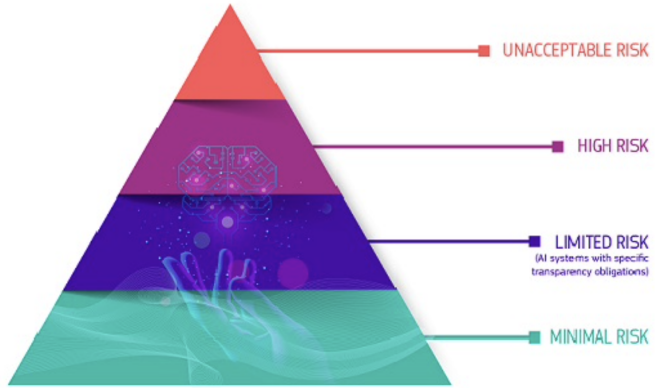
Main pillars

- lawfulness
- ethics
- robustness

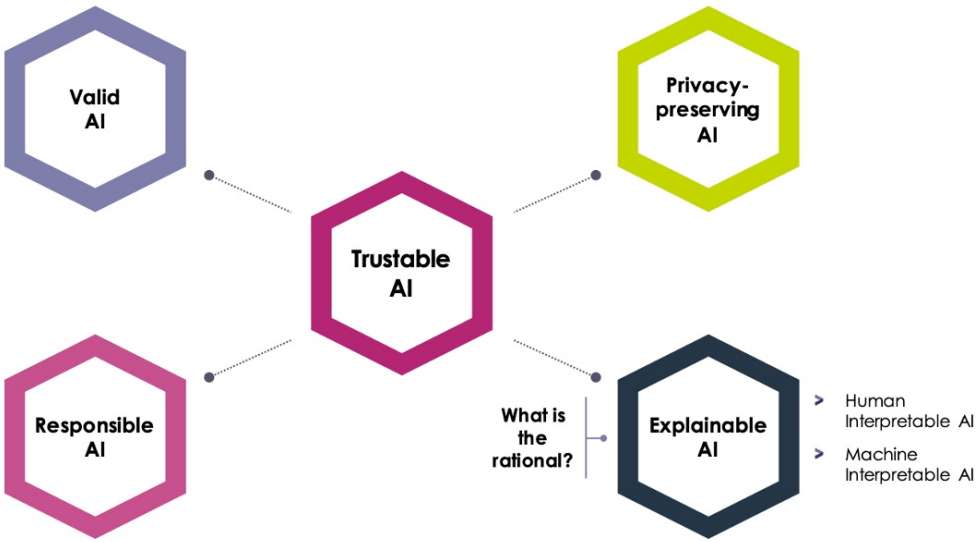
Seven specific requirements – dimensions to be audited – of an AI system:

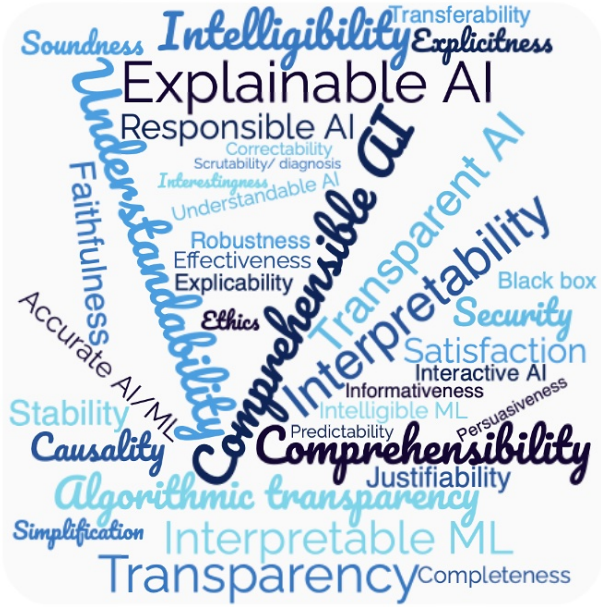
1. human agency and oversight
2. technical robustness and safety
3. privacy and data governance
4. transparency (traceability, explainability)
5. diversity, non-discrimination and *fairness*
6. societal and environmental well-being
7. accountability

AI Act



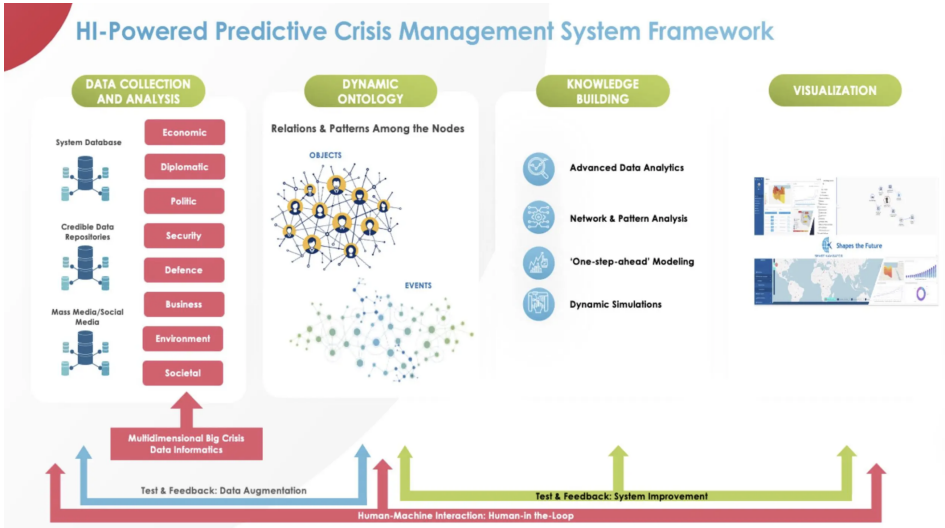
Trustworthy AI: related notions

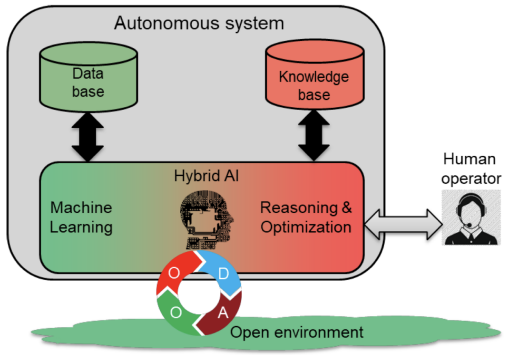


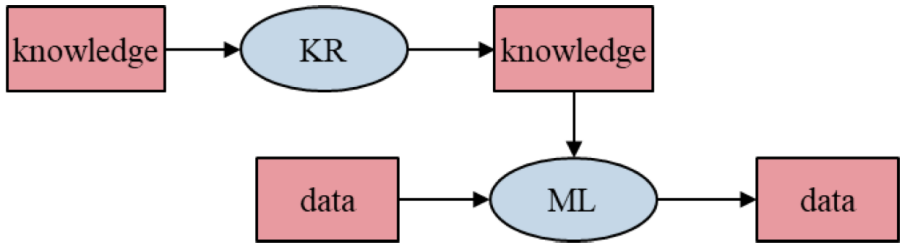


Review of related notions [Vilone and Longo, 2021]

Trustworthy AI: how?







XAI via Integration: Symbolic / Sub-symbolic AI

Symbolic	**Sub-symbolic**
Symbols	Numbers
Logical	Associative
Serial	Parallel
Reasoning	Learning
Localised	Distributed
Rigid and static	Flexible and adaptive
Model abstraction	Fitting to data
Small data	Big data

"What is the added value of symbolic AI for implementing XAI?"

(i) being a declarative paradigm

(ii) working as a tool for knowledge representation

(iii) allowing for different forms of reasoning and inference

(iv) providing a well-founded framework

Gaps, Challenge, Research Directions

- *Educational* aspect of AI practitioners
- *Diversification* is needed beyond existing datasets
- *Assessment metrics* and standardization
- *Experimentation environments* are required to provide an easy playground to test different notions and techniques

References I

[Act, 2021] Act, A. I. (2021).
Proposal for a regulation of the european parliament and the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.
EUR-Lex-52021PC0206.

[European Commission, 2019] European Commission (2019).
Ethics guidelines for trustworthy AI.
Publications Office.

[Leavy et al., 2020] Leavy, S., O'Sullivan, B., and Siapera, E. (2020).
Data, power and bias in artificial intelligence.
arXiv:2008.07341.

[Leavy et al., 2021] Leavy, S., Siapera, E., and O'Sullivan, B. (2021).
Ethical data curation for ai: An approach based on feminist epistemology and critical theories of race.
In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 695–703.

[Vilone and Longo, 2021] Vilone, G. and Longo, L. (2021).
Notions of explainability and evaluation approaches for explainable artificial intelligence.
Information Fusion, 76:89–106.

Trustworthy AI, bias and the role of hybrid intelligence

Roberta Calegari

Alma Mater Studiorum–Università di Bologna, Italy

AI, Data, Robotics Forum #ADRF23
Versailles, Paris

09 November 2023

This work was supported by European Union's Horizon Europe research and innovation programme under grant number 101070363