The relationship between trust in AI and trustworthy machine learning technologies

EHSAN TOREINI,

MHAIRI AITKEN, KOVILA COOPAMOOTOO, KAREN ELLIOTT, CARLOS GONZALEZ ZELAYA, AAD VAN MOORSEL

NEWCASTLE UNIVERSITY

Trust in Computer Science

- Several uses of Trust, e.g., "accepted dependence[1]", Trusted Platform (TPM), etc.
- The definition of trust in society is different.
- This paper:
  - Computer Scientist's perspective to the social science notion of trust and trustworthiness in machine learning.

[1] Avizienis, Algirdas, J-C. Laprie, Brian Randell, and Carl Landwehr. "Basic concepts and taxonomy of dependable and secure computing." *IEEE transactions on dependable and secure computing* 1, no. 1 (2004): 11-33.

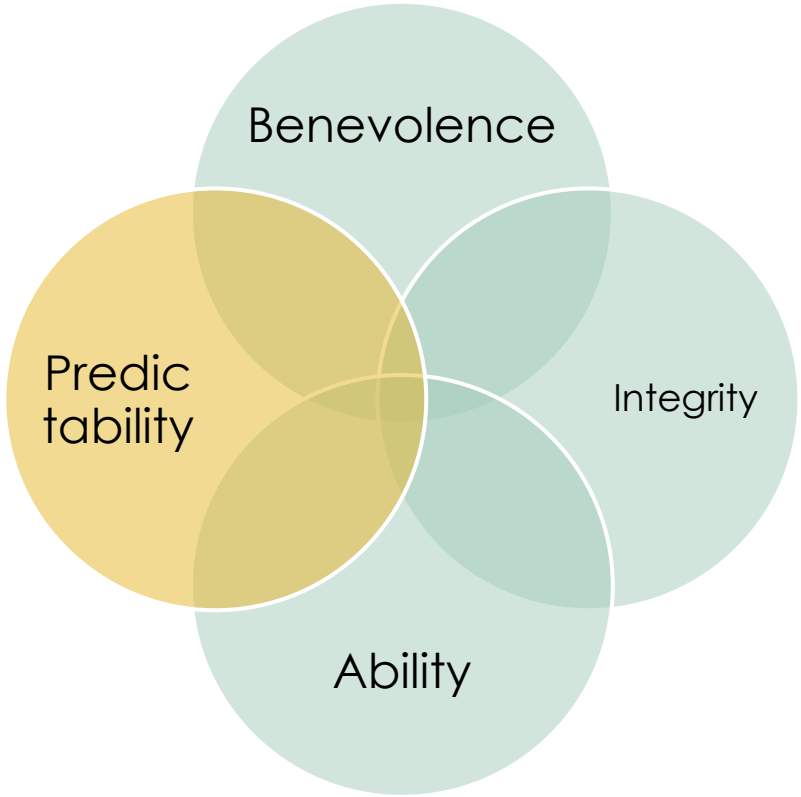
EPSRC
Engineering and Physical Sciences Research Council

The Concept of Trust

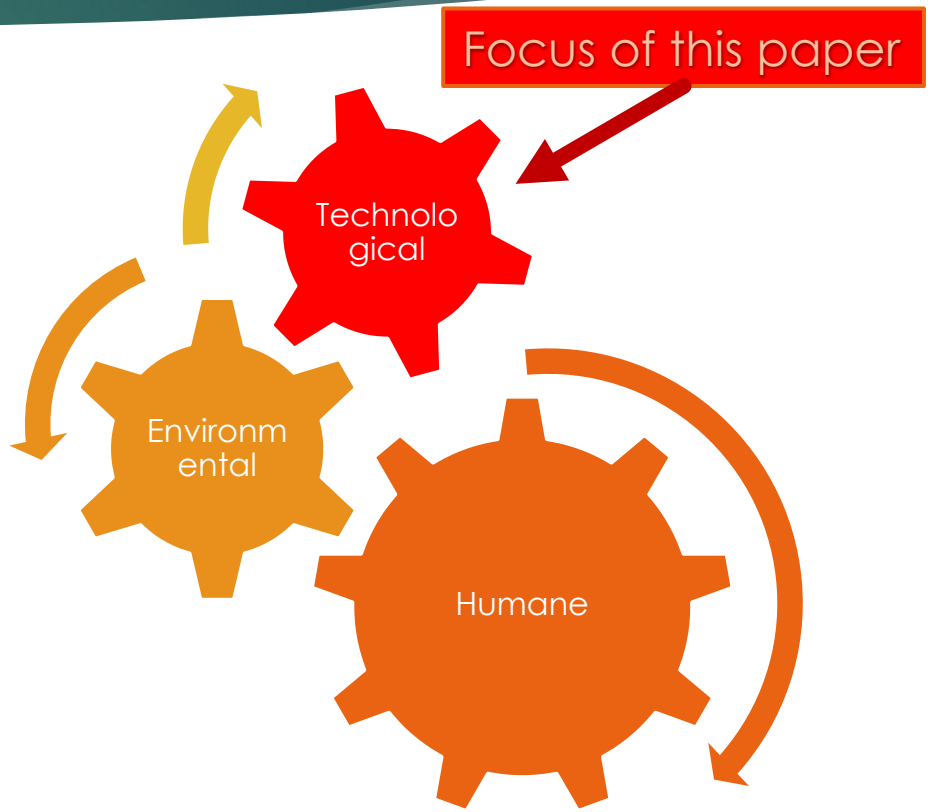
- Everyone has their own personal interpretation for Trust
- In case of machine learning
  - cyber security experts: secure and privacy preserving
  - activists: ethical
  - machine learning experts: accurate and efficient
- Different terminology for trust-related concepts:
  - ethical vs. trustworthy machine learning

EPSRC
Engineering and Physical Sciences Research Council

Trust: In Principle (social sciences)



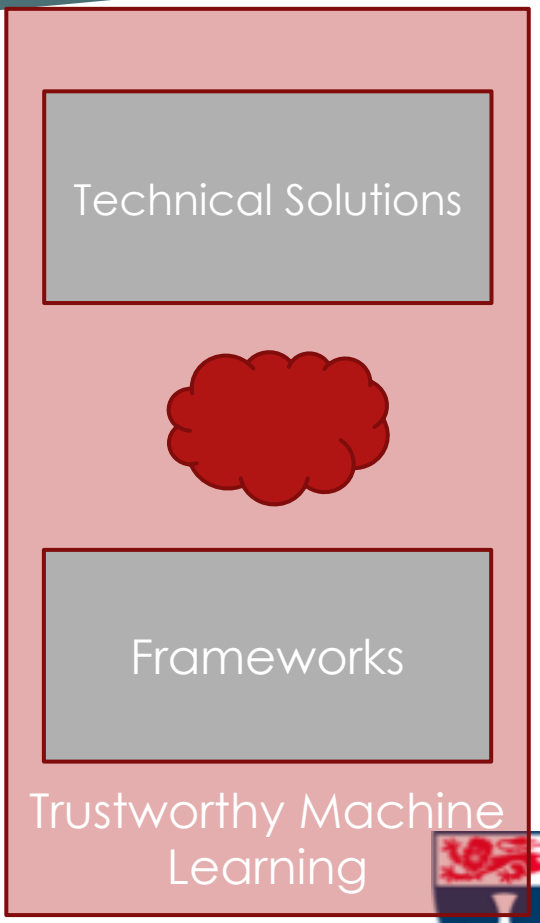
ABI and ABI+ Frameworks (Mayer et al.)



Trust Qualities (Siau et al.)

Trustworthiness: In Reality

- two approaches in the technological requirements for a trustworthy machine learning system:
  - Principled AI frameworks: high level approach
    - Target audience: usually policy makers, governments, industries
  - Technological solutions: low level approach
    - Target audience: usually computer scientists, developers



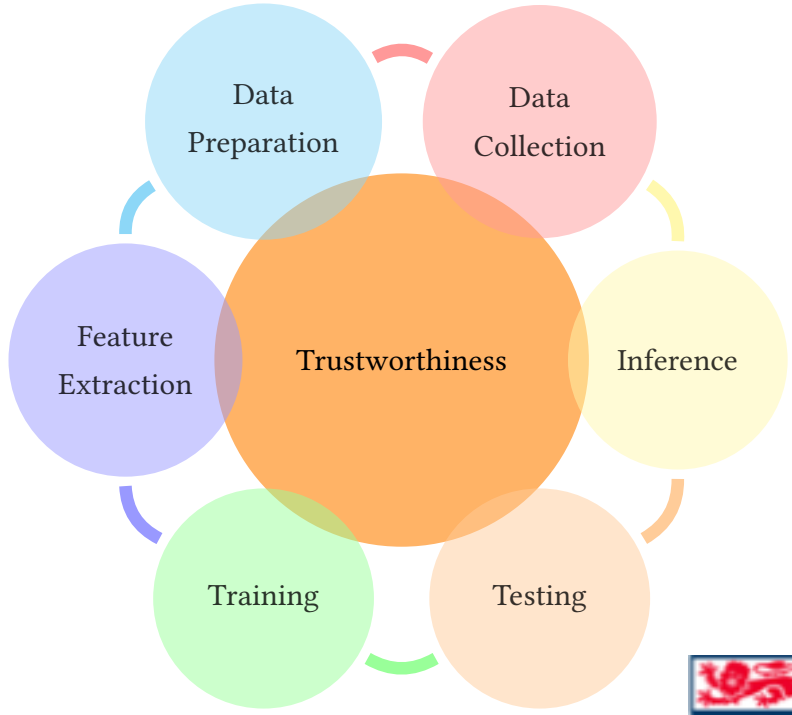
Technological Solutions

- Focused in computer science literature
- We introduce FEAS Technologies as categorisation of Trustworthy solutions:
  - Fairness Technologies
  - Explainability Technologies
  - Auditability Technologies
  - Safety Technologies
- We reviewed 32 frameworks against FEAS technologies:
  - considerable difference in the granularity of the discussions

EPSRC
Engineering and Physical Sciences Research Council

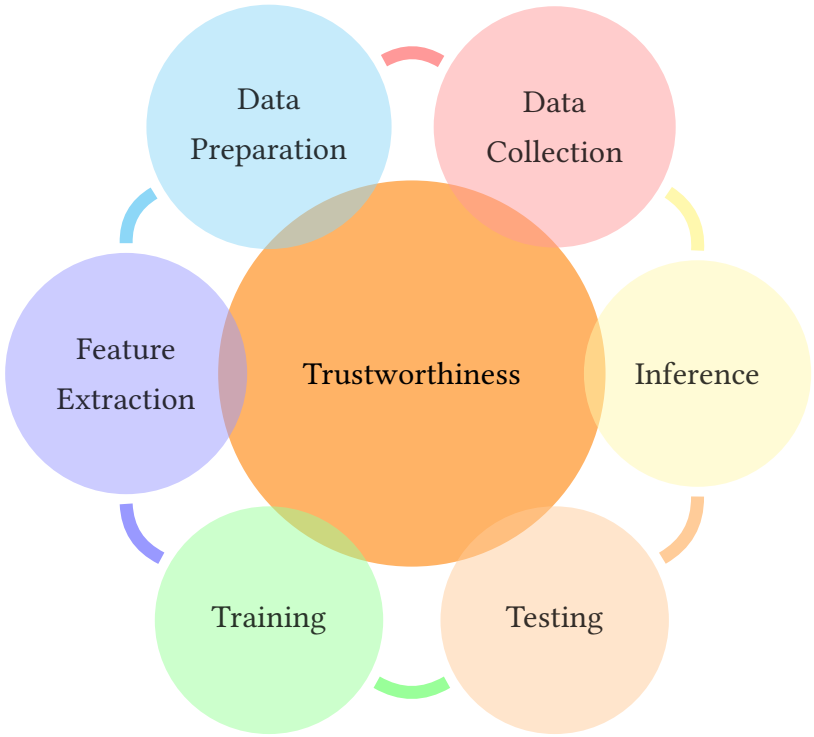
Trust Propagation: FEAS and ML pipeline

- Chain of Trust:
  - Our proposed outlook for implementing FEAS technologies
  - Trust propagates gradually in the machine learning pipeline
- A trust-enhancing solution function in the scope of a few stage.
- Two trust-focused sections:
  - Data Related Trust Solutions
  - Model Related Trust Solutions



Benefits of Considering Chain of Trust

- Stages impact on each other
- Algorithm iterates through the stages during its lifecycle
- Technology decisions in all stages impact others.
- Opportunity to respond to accidents, sudden breakdowns of trust or failures effectively



EPSRC
Engineering and Physical Sciences Research Council

Takeaways

- Trustworthy ML technologies have been subject to various interpretations, e.g., >20 definitions of fairness.
- The trustworthiness technologies do not fully reflect the qualities set by the principled AI frameworks.
- Trustworthiness technologies in different stages of the ML pipeline impact one another: Chain of Trust
- Deeper understanding of how trustworthy ML technologies affect people + societal trust is still needed

EPSRC
Engineering and Physical Sciences Research Council