Can we measure trust?

Agnes DELABORDE
LNE – French national laboratory for metrology and testing
What is an AI that I can trust?

- Safe
- Fair
- Efficient
- Robust
- Resilient
- Reliable
- Traceable
- Auditable
- Ethical
- Controllable
- Accurate
- Predictable
- Explainable
- ...
What is an AI that I can trust?
What is an AI that I can trust?
(example: ALTAI)

Human Agency
Over-sight
Resilience to attack and security
General safety
Accuracy
Reliability, fall-back plans and reproducibility

Human Agency and Autonomy
Technical robustness and safety

Privacy and data governance

Transparency
Privacy
Data governance
Traceability
Explainability

TRUST

Accountability
Societal and environmental well-being

Diversity, non discrimination and fairness

Privacy and data governance

Societal and environmental well-being

Diversity, non discrimination and fairness

Auditability
Risk management
Impact on society and democracy at large
Impact on work and skills
Environmental well-being
Accessibility and universal design
Stakeholder participation
Avoidance of unfair bias

Societal and environmental well-being

Impact on work and skills

General safety

Accuracy

Reliability, fall-back plans and reproducibility

SafeAI 2022 @AAAI-22 – Confiance.ai special session
What is an AI that I can trust?
(example: EASAI)

- TRUST
  - Safety assessment of ML applications
  - Information security for ML applications
  - System safety assessment
  - Information security risk management

- Characterisation of the AI application
  - High-level function/task
  - Concept of operation
  - Functional analysis
  - Classification of the AI application

- Ethics-based assessment
  - Human agency and oversight
  - Technical robustness and safety
  - Privacy and data governance
  - Transparency
  - Diversity, non-discrimination and fairness
  - Accountability

SafeAI 2022 @AAAI-22 – Confiance.ai special session
To what extent can we assess trustworthiness?

- For certain attributes, scores and methods exist:
  - Reliability: Fleiss Kappa score, goodness-of-fit tests, etc.
  - Accuracy: F-measure, precision, recall, etc.

- For other attributes, notions are not fully defined yet:
  - Safety
  - Auditability
  - Absence of bias
  - etc.
Engineer / developer: understand what is important and knows how to verify his/her own work

External inspector: knows precisely what are the checkpoints
Hierarchy and assessment tools

TRUST

Key aspect 1
  Attribute 1.1
  Attribute 1.x

Key aspect 2

Key aspect ...

Key aspect N

Toolbox picture credit: Made from Noun Project
Assessment tools

- **Scores**
  - Result of a computation
  - Result of an observation

- **Methods**
  - Experimentation design

- **Thresholds**
  - Acceptable ranges
  - Acceptable values
Throughout the AI lifecycle

- Problem specification
- Data engineering
- ML algorithm design
- Implementation
- Evaluation and verification
- Model deployment
Taking into account

- Model characteristics
- Algorithm, system and sub-systems
- Operator, impacted/impacting individuals
- Constraints of the context of operation
- Etc.
Can we measure trust?

- Metrological rigor: definition of measure
- Trust encompasses many aspects that are not measurable in themselves
  - Subjective
  - Vague, ill-defined
- Trust is an aggregation of factors (quantitative, qualitative)
  - Good practice from metrology, experimental sciences
  - Multi-criteria assessment
- Assessment of trustworthiness
What are the next steps?

- Defining a trustworthiness score
  - Hierarchy of attributes
  - Rules, methods, scores
  - Overall score of trustworthiness

- Pilot testing on Confiance.ai usecases (critical domain)
  - Usability
  - Relevance

- Transfer to standardization (ISO/IEC JTC1 SC42 Artificial intelligence)