Can we *measure* trust?

Agnes DELABORDE

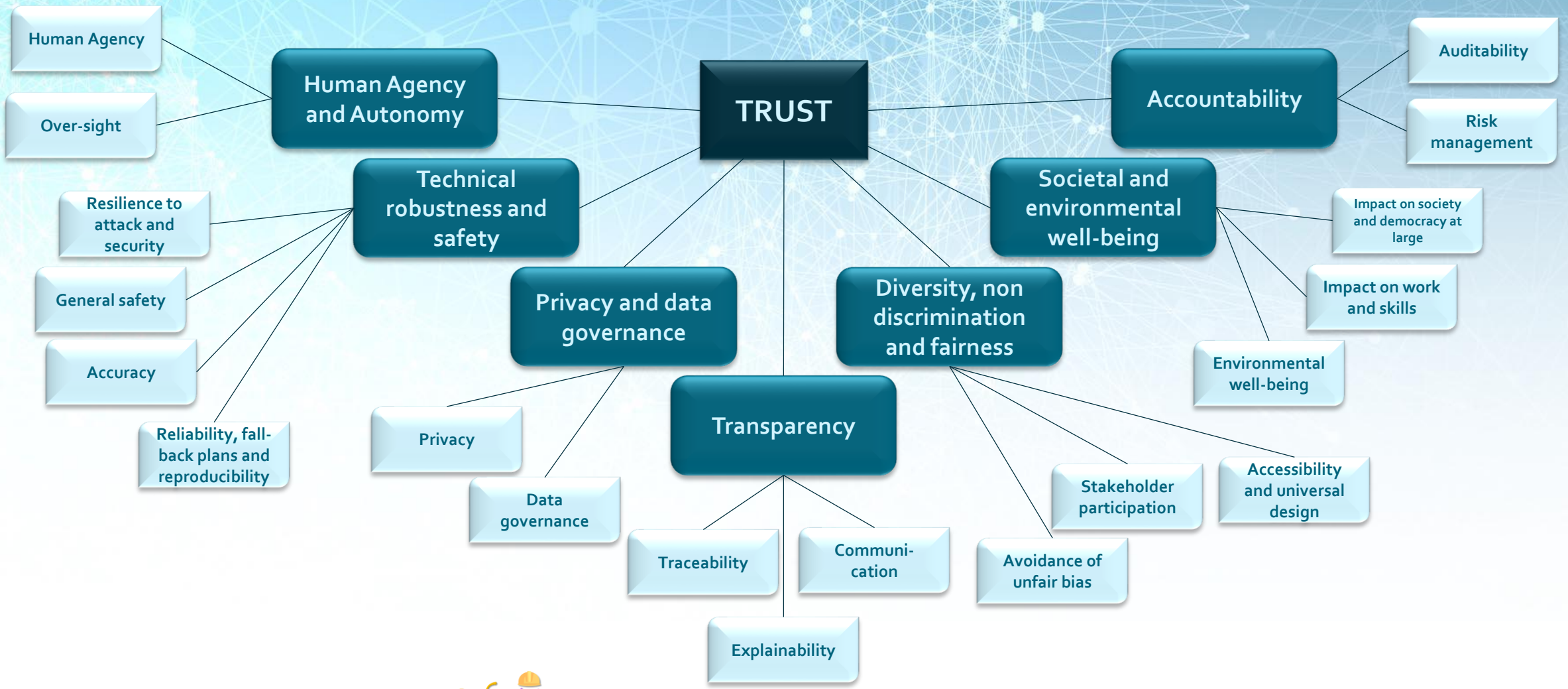
LNE – French national laboratory for metrology and testing

What is an AI that I can trust?

What is an AI that I can trust?

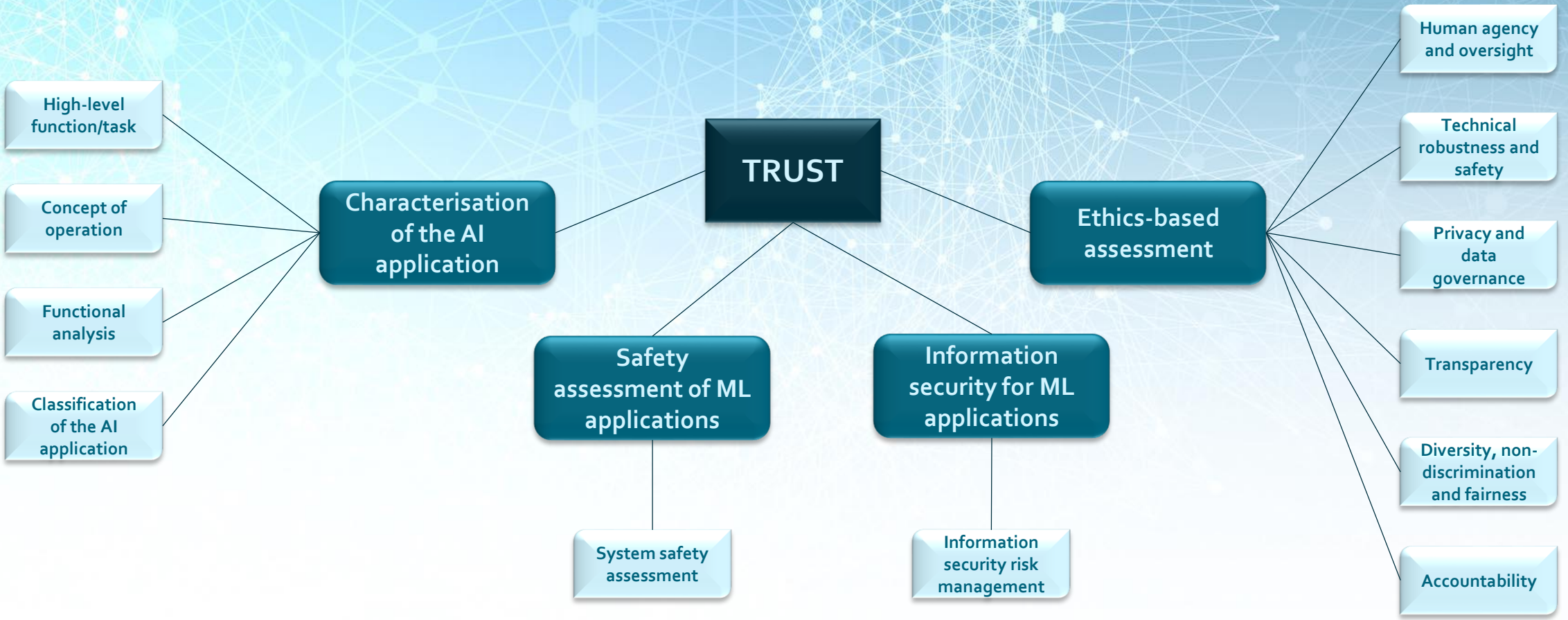
What is an AI that I can trust?

(example: ALTAI)



What is an AI that I can trust?

(example: EASAI)



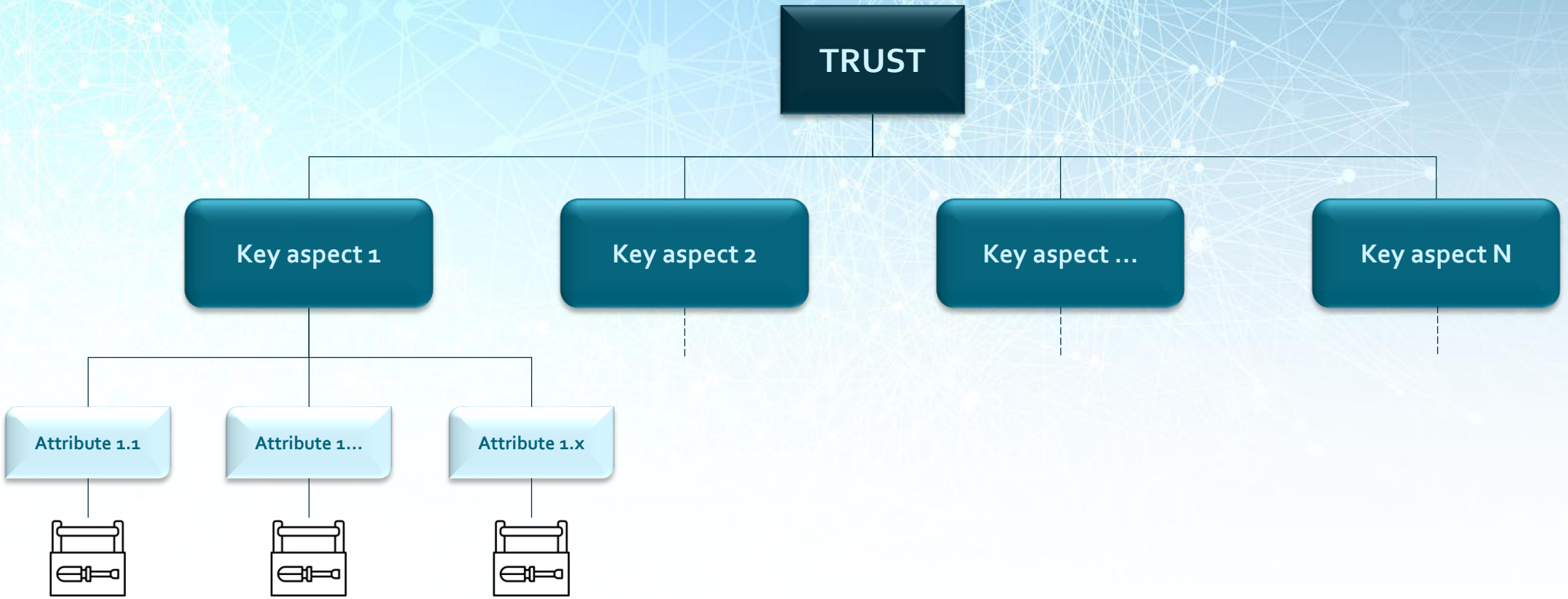
To what extent can we assess trustworthiness?

- **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.

"*not fully defined yet*"

- **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



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

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, qualititative)**
  - 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)**

www.confiance.ai
contact@irt-systemx.fr