Principles of Trusted AI

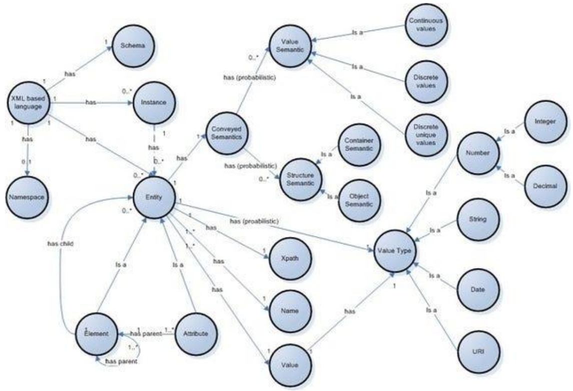
Dr. André Meyer-Vitali, DFKI / CERTAIN

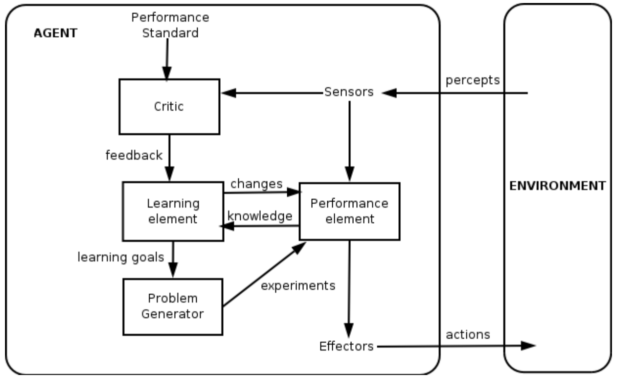
ADR Forum, Versailles – 2023-11-09

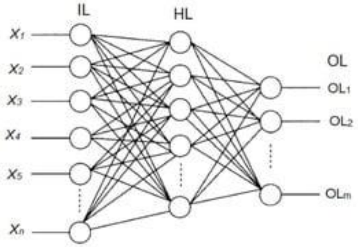
Europäischer Fonds für
regionale Entwicklung
im SAARLAND

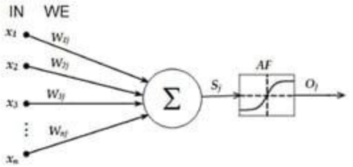
Short History of AI

- Desire to Create Human Beings: Golem, Faust, Frankenstein, Terminator…
- Babbage, Zuse, **Turing**, von Neumann, Wiener, …
- 1956 Dartmouth Workshop: McCarthy, Minsky, Shannon, Newell, Simon
  - **Symbolic Reasoning**: Search, Planning, Logic
  - Expert Systems, Knowledge Engineering, Ontologies
- 1958: McCulloch/Walter Pitts, Rosenblatt: Perceptron
- 1975: **Distributed AI** / Multi-Agent Systems
  - Intelligent Autonomous Agents, Robots
- 1980: John Hopfield, Geoff Hinton, David Rumelhart
  - Artificial **Neural Networks**: Learn from Backpropagation
  - Probabilistic Reasoning
- 2010: **Big Data**
  - **Deep Learning**
  - Pearl: **Causality**
- 2020: **Generative AI** / **Hybrid AI**
  - Large Language Models → Large Multi-Modal Models
  - **Neuro-Symbolic AI**
  - **Hybrid Human-Agent Teams**









We want Trusted AI!

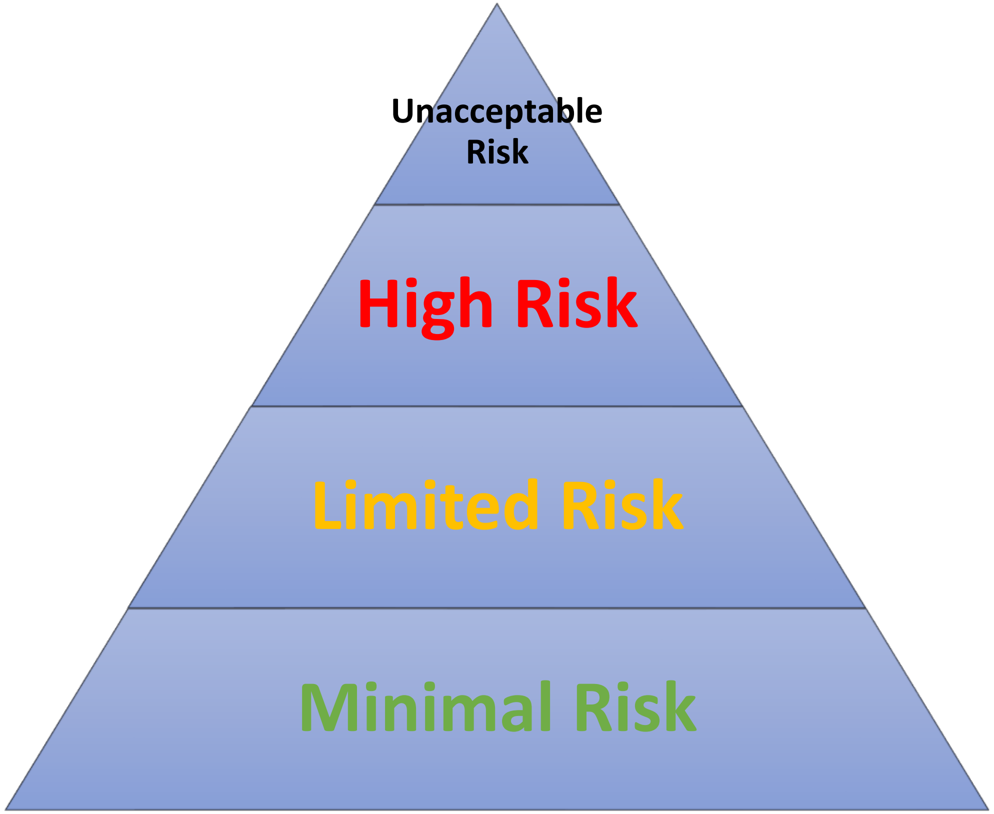
Trust forms the bedrock of our interpersonal relationships and any society.

Without trust we cannot take advantage of the many benefits of AI, but we will also not be able to manage it and avoid negative aspects.

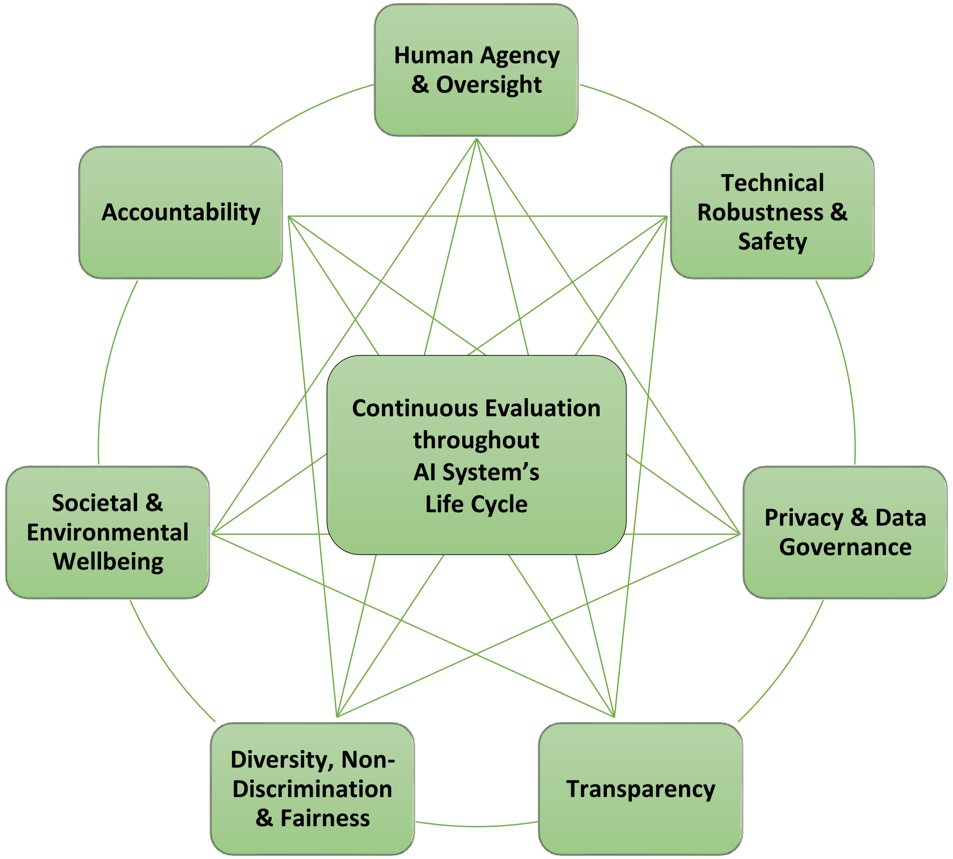
Trusted AI is crucial for critical applications and infrastructure.

Compliance with industry standards and regulations to ensure the safety and reliability of systems.

AI Act & Ethics Guidelines for Trustworthy AI

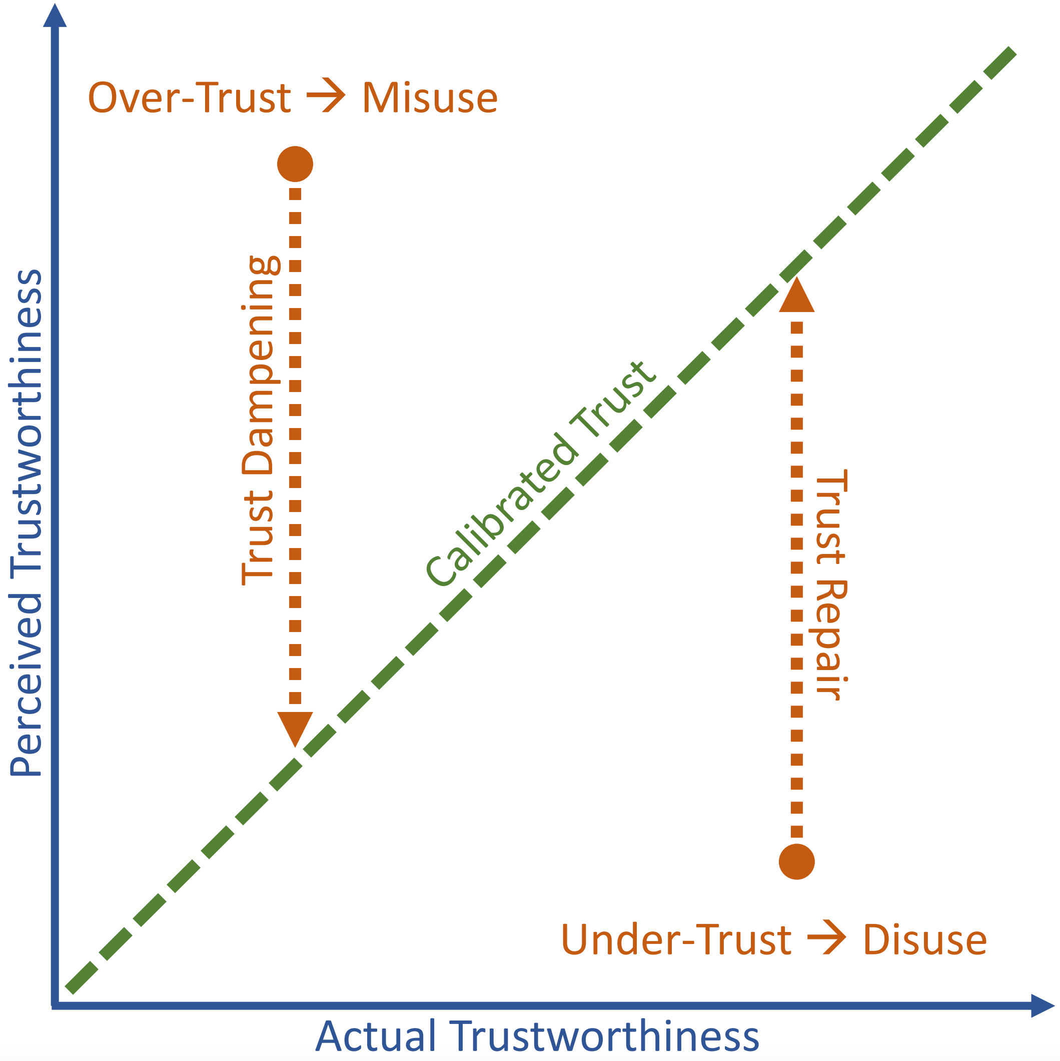


https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai


Human Agency & Oversight
Accountability
Technical Robustness & Safety
Continuous Evaluation throughout AI System's Life Cycle
Societal & Environmental Wellbeing
Privacy & Data Governance
Diversity, Non-Discrimination & Fairness
Transparency

https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html

Trust Calibration



Over-Trust → Misuse

Trust Dampening

Calibrated Trust

Trust Repair

Under-Trust → Disuse

Perceived Trustworthiness

Actual Trustworthiness

de Visser, E.J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics*, *12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

SAARLAND
CERTAIN
Centre for European Research in Trusted Artificial Intelligence

Kick-Off on September 19, 2023, in Saarbrücken.

https://www.certain-trust.eu/

Trusted AI gives Guarantees for…

Functionality and Certifiability

Transparency and Explainability

Fairness and Bias-Awareness

Robustness and Reliability

Safety, Security, Privacy

Responsibility and Accountability

Guarantees for Trusted AI

By Design	By Tools	By Insight	By Interaction
Intrinsic Correctness Deductive Arguments & Proofs (Physical) Laws, Rules, Constraints, Causal Models	Modelling and Simulating the Real World Systematic Testing with Synthetic Data Monitoring, Auditing	Explanations, Reasons Causal Relationships Transparency Visualisation	Human Experience, Influence, Control Human-in-the-Loop Reinforcement Learning from Human Feedback (RLHF) Useable Trust, Trust Calibration

Neuro-Explicit AI Models

Ethics

Standards

Data

Key Aspects of Trusted AI Systems

Models & Explanations

Reliable predictions about system behaviour for insightful and plausible explanations and simulations with generalised models from knowledge and training.

Causality & Grounding

Identification and predictions of cause-effect relationships for informed predictions and anchoring of meaning in real-world context and phenomena.

Modularity & Compositionality

Design of complex systems broken down into comprehensible and manageable parts (functions and features), reliably composed in system architectures.

Human Agency & Oversight

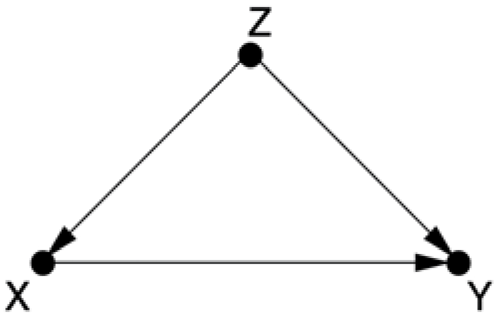
Overview, final decision and responsibility by humans for actions of AI systems, also when delegating tasks to autonomous agents in collaborative teams.

Models & Explanations

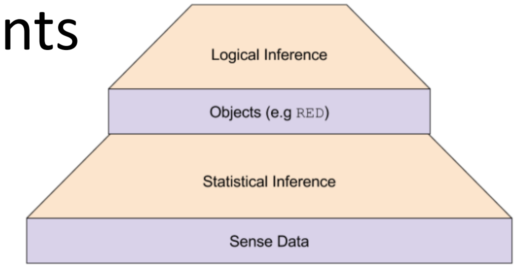
- **Reliable Predictions of the Behaviour of AI Systems**
  - Training Data & Beyond Operational Design Domain (ODD)
  - Out-of-Domain (OOD) Detection and Generalisation
  - Competence Awareness and Adaptation
- **Generation**
  - Created by Experts: Semantic Models
  - Learned from Experience and Data
  - Combinations → Hybrid Models (Neuro-Explicit)
- **Promote Transparency and Explainability**
  - Make AI Systems Understandable and Plausible, Bias-Awareness
- **Simulations, Experiments**

Causality & Grounding

- **Identification and Prediction of Cause-and-Effect Relationships**
  - Asymmetric Relations: Cause → Effect
  - Causal Graphs (DAG) / Structural Causal Models (SCM)
  - Interventions & Counterfactuals
  - Causal Inference & Discovery



- **Anchoring (Grounding) of Meaning in the Real Context**
  - Capture Real-World Phenomena that Data Represents
  - Not only based on Statistical Probabilities
  - Knowledge of Concepts, Contexts, Phenomena and Semantic and Causal Relationships in Reality
  - From Sensation to Representation: Distal Objects


Logical Inference

Objects (e.g RED)

Statistical Inference

Sense Data

Modularity & Structure

- **(Software) Engineering Design Principles: CS → AI**
  - Modular/OO Programming, Design Patterns, Process Engineering
- **Complex Systems Breakdown**
  - Understandable and Manageable Parts (Functions & Features)
  - Reliably Assembled into System Architectures
  - Transitions between successive Components Controlled & Protected
- **Cognitive/Epistemic Models and Languages**
  - Specific, Comprehensible & Verifiable Concepts / Tasks / Tokens
  - Meaning emerges from Structure and Components

Human Agency & Oversight

- **Human Overview, Final Decision and Responsibility**
  - Humans Assess and Approve Actions
  - Accountability
- **Delegation to Autonomous Agents (Software, Robots)**
  - Suitable Task Descriptions
  - Mutual Awareness of Expectations and Intentions
- **Symbiotic Partnership**
  - Hybrid Team Collaboration
  - Complementary Capabilities and Skills
  - Theory of Mind

That's all Folks!

DFKI Newsletter #49 (01/2022)

https://dfki.de/en/web/news-media/news-events/newsletter

TRUSTED AI

BUILDING TRUSTWORTHY AI-BASED SYSTEMS FOR THE FUTURE

© sdecoret, stock.adobe.com

Artificial Intelligence (AI) has emerged as a leading technology in the digital transformation, changing the economy, society, and our lives, while attracting massive investment worldwide. The past decade has been characterized by Deep Learning. Machine learning methods have transformed AI from a niche science to a socially relevant "mega-technology," especially in the fields of image and video analysis as well as in text and language processing.

This new technology is made possible primarily by the latest graphics processors and the availability of vast amounts of data from social media and similar sources.

However, we are coming up against the limits of control over large, highly interconnected, AI-based systems. The complexity of existing AI models is often beyond our understanding, and the methods and processes to ensure safety, reliability, and transparency are lacking. We must overcome these novel and serious limitations or face an inevitable dwindling public and consumer acceptance of AI and dramatic losses in business opportunities and markets. This is clearly visible already in the automotive sector's broad retreat from highly automated driving. AI-based technology is also a key enabler in other German economic sectors – including healthcare, mobility, energy, and the digital industry itself. All of these markets depend on complex and highly connected AI systems designed to support people in decision making and situational analysis.

Despite all the successes, many are not aware that deep learning does not support a real understanding of the problem but only depicts complex statistical relationships. Great disillusionment set in as problems such as insufficient internal representation of meaning (interpretability and transparency), susceptibility to changes in the input signal (robustness), lack of transferability to cases not covered by the data (generalization) and, last but not least, the thirst for big data itself (efficiency, adequacy) became apparent.

Recently, however, a new overall approach to solving these problems is being advanced by the term "Trusted AI." Trusted AI aims to create a new generation of AI systems that guarantee functionality, allowing use even in critical applications. Developers, users, and regulators can rely on performance and reliability even for complex socio-technical systems. Trusted AI is characterized by a high degree of robustness, transparency, fairness, and verifiability where the functionality of existing systems is in no way compromised, but actually enhanced.

Some of the current problems related to a lack of trust in AI systems are a direct result of the massive use of blackbox methods that depend solely on data. Instead, the new AI generation has its foundation built on hybrid AI systems (also known as neuro-symbolic or neuro-explicit). These hybrids do not rely solely on data-driven approaches but on the full range of AI technologies ("All of AI"), which includes symbolic AI methods, search, reasoning, planning, and other operations. "Trust by Design" is achieved through the combination of Machine Learning with symbolic conclusions and the explicit representation of knowledge in hybrid AI systems. Knowledge no longer needs to be machine learned when it is represented by semantic and other explicit models, which can also guide the learning process in a direction that improves generalization, robustness, and interpretability. This hybrid approach is popularly called the third wave of AI ("3rd Wave AI").

The EU's High-Level Expert Group (HLEG) defined principles in the publication "Ethics Guidelines for Trustworthy Artificial Intelligence." AI applications of the future must be legal, robust, and respectful of European ethical principles and values. Legislation obliging these requirements will come into force in the next few years.

The requirements are particularly strict when it comes to applications with significant physical, economic, or social risk. The AI systems used in such applications are assumed to have been validated and certified. Hybrid AI provides exactly the greater transparency that is necessary.

Hybrid AI approaches are studied and applied by the Agents and Simulated Reality research area (ASR) at DFKI, where a newly developed system of possible combinations is helping to assess the advantages and disadvantages in different areas of application. The main research goal of these efforts is Trusted AI. Current research is focused on the area of safety engineering as well as various aspects of validation and certification of AI systems and decision making in human and AI agent teams (Human Empowerment).

Trusted AI implies that trust and reliability can be reformulated as a value proposition. Every economy and society will have to deal with the challenges and threats described above in the near future. Those who manage to push the boundaries of controllability for AI will gain a fundamental competitive advantage. Especially for Germany and Europe, this provides a great opportunity to pool their efforts to strengthen digital competitiveness under a common and visionary goal. "Trusted AI Made in Germany" has the potential to become a globally visible brand that carries Germany's claim of industrial quality leadership into the digital future.

More information

https://tailor-network.eu

Contact

Dr. André Meyer-Vitali
Research Department Agents and Simulated Reality
✉ andre.meyer-vitali@dfki.de
☎ +49 681 85775 7241

TAILOR

© metamorworks, stock.adobe.com