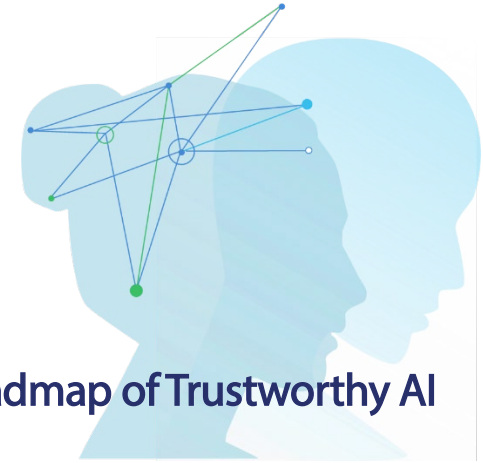
TAILOR

Foundation of Trustworthy AI:
Integrating Learning, Optimisation and Reasoning

2022-07-21
Short extract, V1



Strategic Research and Innovation Roadmap of Trustworthy AI

The Strategic Research and Innovation Roadmap (SRIR) for Trustworthy AI will define the foundations of Trustworthy AI for the years 2022-2030. It aims to boost research on Trustworthy AI by clearly defining the major research challenges.

Objectives

1 PROVIDING GUIDELINES FOR STRENGTHENING AND ENLARGING THE PAN-EUROPEAN NETWORK OF RESEARCH EXCELLENCE CENTRES ON THE FOUNDATIONS OF TRUSTWORTHY AI

2 DEFINING PATHS FOR ADVANCING THE SCIENTIFIC FOUNDATIONS FOR TRUSTWORTHY AI AND TRANSLATING THEM INTO TECHNICAL REQUIREMENTS TO BE ADOPTED BROADLY BY INDUSTRY.

3 IDENTIFYING DIRECTIONS FOR FOSTERING COLLABORATIONS BETWEEN ACADEMIC, INDUSTRIAL, GOVERNMENTAL, AND COMMUNITY STAKEHOLDERS ON THE FOUNDATIONS OF TRUSTWORTHY AI

Recommendations

Measure and assess Trustworthy AI dimensions

Short term

- Develop methods for measuring and evaluating the trustworthiness of AI systems.

Long term

- Develop tools for continuously auditing and adapting Trustworthy AI systems: monitoring, dynamically identifying issues, and mitigating them.

TAILOR is an ICT-48 Network of AI Research Excellence Centers funded by EU Horizon 2020 research and innovation programme under grant agreement No 952215

Scientific challenges

Short term

- Develop human interpretable formalisms to enable synergistic collaboration between humans and machines with regards to the criteria of being explainable, safe, robust, fair, accountable; and develop standards and metrics to quantify the grade to which these criteria are satisfied.
- Develop methods for integrating model-based and data-driven approaches to autonomous acting.
- Develop a broad range of AutoAI benchmarks to facilitate development and critical assessment of AutoAI techniques and systems.
- Expand current AutoAI techniques to better meet the demands of real-world applications, including multiple interacting design objectives (with aspects of trustworthiness), scalability, scope and ease of use.
- Develop integrated representations and frameworks for learning, reasoning and optimisation based on probability, logic, neural networks, ontologies, knowledge graphs and constraints.

Long term

- Develop the science, techniques and tools for adjustable autonomy for autonomous AI agents. In particular, equip autonomous agents with the ability to understand when certain decisions that it could take on its own are questionable or unethical, and human supervision should be required.
- Develop a computational theory of mind that considers mental attitudes such as beliefs, knowledge, goals, intentions, capabilities, emotions, and integrates them in a computational effective fashion into autonomous acting.
- Enable the broad, safe, and efficient use of AutoAI techniques across all sectors of industry and society, especially in contexts where limited AI expertise is available.
- Develop a unifying theory and framework of learning, reasoning and optimisation that that bridges the gap between the data- and knowledge-driven and the symbolic and subsymbolic approaches in AI.

Innovation

Short term

- Develop generic operational models of hybrid approaches allowing their reuse in various domains and propose metrics/benchmarks for validating these models.
- Consider that transparency (incl. explainability) targets different kinds of users: developers, domain experts, regulators, “users” (citizens, patients, etc.).

Long term

- Implement Trust by Design: Enable the design and verification of trusted AI systems according to appropriate legal, social and technical criteria and aspects, focusing in particular on critical and risky applications.

Read more: www.tailor-network.eu