Contents

1 Foundations of Trustworthy AI

> Content from TAILOR deliverable report D9.6
>
> The content of this page is currently a re-formatted copy from the Deliverable 9.6 PhD Curriculum Report.

This topic covers the dimensions of Trustworthy AI: (i) Explainability, (ii) Safety, (iii) Fairness, (iv) Accountability and Reproducibility, (v) Privacy, and (vi) Sustainability.

1.1 Content/Knowledge

Students should be able to understand/describe current discourse on the following questions:

- How can we guarantee user trust in AI systems through explanation? How to formulate **explanations as Machine-Human conversation** depending on context and user expertise?
- How to bridge the gap from safety engineering, formal methods, verification as well as validation to the way AI systems are built, used, and reinforced?
- How can we build algorithms that respect **fairness constraints** by design through understanding causal influences among variables for dealing with bias-related issues?
- How to uncover **accountability** gaps w.r.t. the attribution of AI-related harming of humans?
- Can we guarantee **privacy** while preserving the desired utility functions?
- Is there any chance to reduce energy consumption for a more **sustainable AI** and how can AI contribute to solving some of the big sustainability challenges that face humanity today (e.g. climate change)?
- How to deal with properties and tradeoffs among multiple dimensions? For instance, accuracy vs. fairness, privacy vs. transparency, convenience vs. dignity, personalization vs. solidarity, efficiency vs. safety and sustainability.

1.2 Methodological/Skills

Students should be able to:

- apply their critical and analytical faculties on specific case studies, in order to argue about the need and content of AI trustworthiness issues.

1.3 Transferrable/Application

Students should be able to:

- Work effectively with others in an interdisciplinary and/or international team.
- Clearly and succinctly communicate their ideas to technical and non-technical audiences.