TAILOR

4th TAILOR Conference

Trustworthy AI from lab to market

4-5 June 2024 in Lisbon, Portugal

A Card-based Agentic Framework for Supporting Trustworthy AI

Mattheos Fikardos, Katerina Lepenioti, Dimitris Apostolou, Gregoris Mentzas

Information Management Unit, ICCS, School of Electrical and Computer Engineering
National Technical University of Athens, Greece.

Abstract

The rapid advancements in AI have triggered the need for Trustworthy AI (TAI), which encompasses various definitions, perspectives, and technological approaches aimed at ensuring that AI is reliable and trusted by humans. Already organizations, governments, and academia have produced regulations and frameworks around TAI (e.g. the EU AI Act, the NIST Risk Management Framework), but a gap still exists between those ethical and legal guidelines and their practical applicability by companies. Despite the plethora of methods and algorithms that assess or enhance AI trustworthiness, these remain fragmented, each targeting a specific aspect of trustworthiness (e.g. accuracy, transparency, fairness, explainability). This gap increases the need for a holistic approach that would provide support to companies developing and deploying trustworthy AI systems.

Our work primarily tries to address this gap with a methodological framework that unifies and entangles the AI development lifecycle with the trustworthiness requirements and integrates them within a software solution. We define the lifecycle phases of an AI system, from its design and development through deployment and monitoring, as well as the trustworthiness phases, following a risk management approach where AI risks are identified, assessed, and mitigated. In order to record and structure information related to TAI within our framework, we use and extend the "card-based" approach. We collect information through data cards, model cards and use case cards.

We also propose the use of "methods cards", which structure technical information about available algorithms, methods and software toolkits that assess and enhance AI trustworthiness. For the development of our software framework, we follow a neuro-symbolic agentic design pattern in which different roles involved in the TAI assessment process can be instantiated. This is done by enabling an LLM to be prompted and guided by navigating knowledge graphs, which have been derived from the already recorded cards.

Currently, we have developed an initial prototype which, at this stage, focuses on two TAI dimensions: fairness and robustness. We have recorded more than 20 methods for each for these dimensions and developed the corresponding cards. Our aim is to further work on additional TAI dimensions and evaluate our framework in three distinct cases: disinformation and fake news detection; cancer risk identification and assessment; and management of disruptive events in ports.

Acknowledgement: This work is funded by the EU Horizon Europe programme CL4-2022-HUMAN-02-01 under the project THEMIS 5.0 (grant agreement No.101121042) and by UK Research and innovation under the UK governments Horizon funding guarantee. The work presented here reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.