The Design of Trustworthy AI System: a Deep Look into the Transparency of Data, Models, and Decisions

Sara Colantonio

The Design of Trustworthy AI System: a Deep Look into the Transparency of Data, Models, and Decisions

Sara Colantonio, ISTI-CNR, Pisa, Italy

***Abstract*— Transparency is a cornerstone of trustworthy AI, a key mandate of the European approach to AI and a cross-sectional principle of FUTURE-AI guidelines in health imaging. In this talk, we will critically overview the strategies to guarantee an appropriate level of transparency, mainly in terms of data provenance and usage, AI system development and functioning interpretation. Current approaches, hindrances, and opportunities will be highlighted and discussed, along with the strategies adopted in the EU H2020 ProCAncer-I project.**

I. INTRODUCTION

Realizing the full potential and benefit of Artificial Intelligence (AI) solutions in high-stake domains, such as medical imaging diagnostics, requires high-quality scientific foundations, technical robustness, and responsible development. This vision is at the core of the European approach to AI [1], which promotes excellence and trust as the main drivers of a beneficial impact of AI. Undeniably, only those applications that guarantee reliability, stakeholders' trust and acceptance, and total patients' safety can be expected to have a real impact and uptake in clinical practices. A key pillar of trustworthiness is transparency, which entails to document the entire life-cycle of an AI system as well as the underlying principles of its functioning.

Making an AI system *transparent by design* is key to avoid any grey area in its functioning and use by decision makers in clinical practice. Therefore, it is an overarching principle of the FUTURE-AI guidelines [2], notably touching upon the Traceability, Explainability and Usability principles. Transparency also ensures that the AI system is reproducible and auditable by design, thus laying the bases for accountability and liability. In this work, we illustrate the core dimensions of transparency in medical imaging and discuss current approaches, along with their limits and hindrances, thus overviewing the solutions under development in the EU H2020 ProCAncer-I project.

II. THE CORE DIMENSIONS OF TRANSPARENCY

AI solutions for medical imaging mainly rely on data-driven methods able to process high-volume multimodal data. Transparency in this field is a multifaceted concept that requires best practices and technical measures along the entire value-chain of AI, from data collection, to system development and deployment. The main dimensions are:

- *data transparency*: via a complete documentation of data *provenance* (ownership, acquisition modalities, reference clinical standards, curation, storage, and processing)
- *model transparency*: via a complete and standardized report on model purpose, development choices, performance, failures
- *decision transparency*: via the provision to clinicians of meaningful and actionable explanations about the logic behind any AI classification or prediction results. Explanations should be sound and respond to users' needs, as this is essential to make them feel in control and really empowered by the AI system.

Currently, standards have been proposed mainly for data provenance, whilst model transparency is still in its infancy, with MLOps frameworks featuring part of the required functionalities. Around decision transparency the discipline named eXplainable AI (XAI) has recently flourished, though being an old quest in decision-making [3], [4]. XAI is key to realize profitable human-AI teaming, but it requires techniques belonging to multiple domains (see Figure 1.).

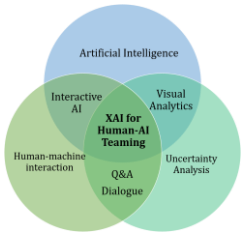

Figure 1. Profitable XAI emerges at the crossroads of diverse disciplines.

III. DISCUSSION & CONCLUSION

In ProCAncer-I, we are working to ensure transparency along all its dimensions, by adopting data provenance standards, defining the so-called *Model Passport* for model documentation and devising innovative strategies to explainability. In particular, we are working together with clinicians to define the best explanations modalities and the metrics to quantify their understandability by end-users. The goal is to convey the necessary information to enable users to *team* with AI for powered and informed decisions.

This work has been partially funded by the ProCancer-I European Union's H2020 programme under Grant Agreement No. 952159

REFERENCES

[1] EC's Comm "*Artificial Intelligence for Europe*", COM(2018)237
[2] K. Lekadir, et al. *FUTURE-AI: Guiding Principles*, arXiv, 2021
[3] F. Chiarugi et al. *Biomedical signal and image processing for decision support in heart failure*. AIM, pp. 38-51, Elsevier, 2008
[4] S. Colantonio et al. *An approach to decision support in heart failure*. In Proc. of SWAP 2007