Understanding the Risks of AI-Systems: A Risk Management Approach

Background

As stated in multiple guidelines and upcoming european regulations, transparency and explainability are a requirement for ethical by design AI-powered technologies. Explainability of AI can be understood from various angles: interpretability for the developers, and understandability for the users. It is this second side that interests us in this research, as a step towards an accountability framework for AI-systems.

Expected goals

Drawing upon past literature, and survey designed data collection, this thesis will aim at investigating individual characteristics linked to AI-systems risks identification and understandability. To do so, adequate definitions and vocabulary linked to specific terms and risks will have to be identified, as well as adequate visual and sound designs. Additionally, the acceptable trade-offs of opacity and explainability in the context of AI-Systems from a user perspective will be evaluated.

Research Questions

- Which are risks linked to AI-Systems identifyable by the users?
- How to explain the different risks present in the use of different AI-Systems to users depending on individual characteristics?
- Which design (e.g., vocabulary, explanations, visuals) are to be used with the users depending on individual characteristics to insure understandability?

Recommended literature

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018, April). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-18).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689-707.

Xu, W. (2019). Toward human-centered AI: a perspective from human-computer interaction. *Interactions*, *26*(4), 42-46.

Details

Supervisor:	Auxane Boch
Starting date:	as soon as possible

Contact

If you are interested, please contact Auxane Boch (**auxane.boch@tum.de**).

We are looking forward to your application!