Computational Responsibility for Trustworthy Citizen-Centric AI

Vahid Yazdanpanah, Jennifer Williams and Sebastian Stein

University of Southampton

As AI systems become increasingly integrated into our lives, it is essential to ensure that they are reliable and trustworthy, especially from the perspective of citizens. One way to achieve this is through the use of computational methods for reasoning about different forms of *responsibility* (such as accountability for delivering tasks and liability for harmful behaviour). In our research, we explore how such methods can support the development and trustworthy deployment of citizen-centric AI systems [3, 7].

To achieve this goal, we propose using both backward-looking and forward-looking responsibility models [6]. Backward-looking responsibility models involve analysing data post-hoc (e.g., using data-mining techniques) to reason about the historical behaviour of AI systems, identifying potentially liable candidates for harm, and quantifying the extent of agents' responsibility [5]. This approach can serve as a valuable decision support tool for the judiciary system as it allows decision-makers to focus on a smaller number of scenarios and enables them to navigate complex structures and causal chains. For instance, as discussed in [1], suppose an autonomous vehicle causes an accident. In that case, backward-looking responsibility models can help determine if the vehicle was at fault, whether it was the software or the hardware component that malfunctioned, or if the manufacturer failed to put in place adequate harm-avoidance mechanisms before deploying the system. This type of analysis can assist in determining legal and ethical responsibilities for such incidents and enable monitoring of system design and deployment.

On the other hand, forward-looking responsibility models focus on using formal methods and system verification techniques to (prospectively) evaluate how a system may perform in interaction with its users and the environment, e.g., in a simulated city environment, and to see whether the designed tasks are deliverable or if some harmful states are likely to occur. By considering the needs and preferences of different groups of citizens, such models can help developers ensure that their systems are tailored to meet the specific needs of these groups and that harm (in view of citizens) is maximally avoided. This approach can promote the development of more reliable and trustworthy AI systems better suited to the needs of citizens. Consider the case of an audio-based AI assistant designed to help users with their needs. While such AI assistants hold promise, they may also pose a risk of breaching users' privacy, which is a user-specific concern [4]. To mitigate the risk of breaching privacy, a forward-looking responsibility model can verify that the system performs its tasks securely in a privacy-aware manner and meets the required standards. This is by verifying if components are allocated a task they can deliver safely, that the combination of tasks does not lead to any breach of privacy and meets the expectations, and that for any potential breach, the accountable parties are clear. Forward-looking responsibility models can thus prevent harmful outcomes, improve system reliability, and enhance trust in AI systems among citizens.

In conclusion, both backward-looking and forward-looking responsibility models can promote the development of citizen-centric AI systems that are legally compliant, ethically responsible, and tailored to the needs of citizens. These models can serve as valuable decision support tools for the judiciary system and help ensure that AI systems are reliable, trustworthy, and beneficial to all stakeholders, enabling an effective embedding of AI technologies into society [2].

Acknowledgement: This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through a Turing AI Fellowship (EP/V022067/1) on Citizen-Centric AI Systems (`https://ccais.ac.uk/`).

References

[1] Mehdi Dastani and Vahid Yazdanpanah. Responsibility of AI systems. *AI & SOCIETY*, pages 1–10, 2022.

[2] Sarvapali D Ramchurn, Sebastian Stein, and Nicholas R Jennings. Trustworthy human-AI partnerships. *Iscience*, 24(8):102891, 2021.

[3] Sebastian Stein and Vahid Yazdanpanah. Citizen-centric multiagent systems. In *Proceedings of The 22nd International Conference on Autonomous Agents and Multiagent Systems*, 2023.

[4] Jennifer Williams, Vahid Yazdanpanah, and Sebastian Stein. Safe audio AI services in smart buildings. In *International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 266–269, 2022.

[5] Vahid Yazdanpanah and Mehdi Dastani. Quantified degrees of group responsibility. In *Coordination, Organizations, Institutions, and Norms in Agent Systems*, pages 418–436. Springer, 2016.

[6] Vahid Yazdanpanah, Enrico H Gerding, Sebastian Stein, Corina Cirstea, m.c. Schraefel, Timothy J Norman, and Nicholas R Jennings. Different forms of responsibility in multiagent systems: Sociotechnical characteristics and requirements. *IEEE Internet Computing*, 25(6):15–22, 2021.

[7] Vahid Yazdanpanah, Enrico H Gerding, Sebastian Stein, Mehdi Dastani, Catholijn M Jonker, Timothy J Norman, and Sarvapali D Ramchurn. Reasoning about responsibility in autonomous systems: challenges and opportunities. *AI & SOCIETY*, pages 1–12, 2022.