Task-Oriented Accountability in Autonomous Systems

Vahid Yazdanpanah[1], Sebastian Stein[1], Enrico H. Gerding[1], and Nicholas R. Jennings[2]

[1]University of Southampton
[2]Imperial College London

September 2021

In Artificial Intelligence (AI) systems, a key problem is to determine the group of agents that are accountable for delivering a task and, in case of failure, the extent to which each group member is partially accountable. In this context, accountability is understood as being responsible for failing to deliver a task that a team was allocated and able to fulfil. This is, on one hand, about agents' accountability as collaborative teams and, on the other hand, their individual degree of accountability in a team. Developing verifiable methods to address this problem is key for designing trustworthy autonomous systems and ensuring their safe and effective integration with other operational systems in society. Using degrees of accountability, one can trace back a failure to AI components and prioritise how to invest resources on fixing faulty components.

In this talk, we report on a line of research [2] on the application of formal methods and modal logics for reasoning about accountability in multiagent systems and focus on answering *"Who is accountable for an unfulfilled task in multiagent teams: when, why, and to what extent?"*. In addition, we elaborate on open problems [1], link to ensuring safety in application domains such as Connected and Autonomous Vehicles (CAVs), and highlight the potentials of formal accountability reasoning in design and development of trustworthy AI systems.

Acknowledgements. This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the Trustworthy Autonomous Systems Hub (EP/V00784X/1), the platform grant entitled "AutoTrust: Designing a Human-Centred Trusted, Secure, Intelligent and Usable Internet of Vehicles" (EP/R029563/1), and the Turing AI Fellowship on Citizen-Centric AI Systems (EP/V022067/1).

References

[1] Vahid Yazdanpanah, Enrico H. Gerding, Sebastian Stein, Mehdi Dastani, Catholijn M. Jonker, and Timothy J. Norman. Responsibility research for trustworthy autonomous systems. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems-AAMAS-2021*, page 57–62, 2021.

[2] Vahid Yazdanpanah, Sebastian Stein, Enrico H. Gerding, and Nicholas R. Jennings. Applying strategic reasoning for accountability ascription in multiagent teams. In *The IJCAI-21 Workshop on Artificial Intelligence Safety (AISafety 2021)*, August 2021.