Explaining Themselves and Making Friends: Towards Formalising the Sociability of Autonomous Agents

Maike Schwammberger[1], Akhila Bairy[1], and Barbara Bruno[1]

Karlsruhe Institute of Technology, Karlsruhe 76131, Germany
{schwammberger,akhila.bairy,barbara.bruno}@kit.edu

Abstract. While autonomous systems are integrated into more and more close-to-human application domains, we investigate sociability as a necessary extra-functional system property to ensure their integrability into diverse societies. To enable formalisation and formal validation of sociability of autonomous systems, we derive requirements for social rules from interdisciplinary sources. We further discuss explainability as a tool for understanding social actions of autonomous systems.

1 Motivation

Automated and embodied software-intensive systems are playing a more and more central role in our everyday lives. From autonomous vacuum cleaners to intelligent factory robots and automated cars: Ensuring the safety and reliability of these systems is and must continue to be a central research focus. For humans to accept autonomous systems among them in increasing numbers of application domains, an extra-functional system property comes to the fore: *sociability*. Our meaning of the term sociability encompasses an agent's capability of appropriate social behaviour and traits like approachability, responsiveness, and adaptability in social contexts.

Our talk focuses on autonomous agents, a terminology comprising the software entity steering an automated system. We *identify the need to formalise and validate the sociability* of autonomous agents to ultimately ensure their integrability into dynamic, close-to-human, real-life application contexts. We postulate that formalised *social rules* must be *embedded into existing rule books* for autonomous agents. Through machine-readable social rules, an autonomous agent is enabled to reason about context-dependent social behaviour autonomously. Further on, formalised rule sets pave the way towards formal guarantees and proofs of social agent behaviour.

2 Key Challenges and Connection to the FMIAI Track

A key challenge for reasoning about sociability lies in its inherently non-formal and vague nature: What is perceived as acceptable social behaviour depends on a variety of societal and cultural structures and contexts. While formal methods

provide the tools to validate sociability and are necessary to even begin with integrating reasoning about sociability into autonomous agents, formalising human behaviour and social norms is an unsolved problem. To tackle this challenge, we suggest that logic- and rule-based research directions ("symbolic Artificial Intelligence (AI)") must join forces with human-robot interaction (HRI) and empirical, learning-enabled and probabilistic systems' research ("sub-symbolic AI"). As HRI investigates the interaction of embodied autonomous systems with humans, learning from HRI research findings is an ideal starting point for our endeavour to formalise sociability. Furthermore, approaches on learning of social behaviour from humans will provide more insights into characteristics of social rules [Le18]. Besides learning sociability from humans, it will be interesting to investigate learning of different societal contexts for formalising context-dependent social rules.

3 Our Contribution

While formalising social norms for autonomous agents is an underexplored research area in formal methods research, social behaviour is a key research focus in learning-enabled HRI research [BB24; BBV24]. We first review and analyse notions of sociability that are suggested in HRI research [Da07; PA10]. From this structured analysis, we derive requirements for social rules for autonomous agents and investigate suitable formalisation means. For this, we build upon our previous work on formalising traffic rules for autonomous traffic agents [RS23; Sc18; Sc25]. To enable agents to understand human actions and reactions, we are taking approaches on formalising human behaviour through cognitive models and attention models into account [FBH23; Wi15]. Finally, we acknowledge that communication plays a crucial role in sociability [Bl19]. We thus investigate self-explainability as a crucial enabler of sociability. Through an explanation, a human can understand the social intention of an autonomous agent, even in opaque and dynamically shifting application contexts [Ba22; BF23; Sced].

Our ultimate goal is to build an interdisciplinary roadmap that will allow us to address the sketched challenges and ultimately aims to contribute to the design of safer, more reliable and more acceptable autonomous systems.

References

[Ba22] Bairy, A.: Modeling Explanations in Autonomous Vehicles. In (ter Beek, M. H.; Monahan, R., eds.): Integrated Formal Methods. Springer International Publishing, Cham, pp. 347–351, 2022, ISBN: 978-3-031-07727-2.

[BB24] Burkart, D.; Bruno, B.: Human-like Social Learning for Social Robots: A Systematic Review. In: 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN). Pp. 1–7, 2024, DOI: 10.1109/RO-MAN60168.2024.10731225.

[BBV24] Bied, M.; Bruno, B.; Vinel, A.: Autonomous Vehicles as Social Agents: Vehicle to Pedestrian Communication from V2X, eHMI and HRI Perspectives. In: 2024 20th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). Pp. 86–91, 2024, DOI: 10.1109/WiMob61911.2024.10770473.

[BF23] Bairy, A.; Fränzle, M.: Optimal Explanation Generation Using Attention Distribution Model. In: Human Interaction and Emerging Technologies (IHIET 2023). Vol. 70, AHFE Open Access, pp. 41–49, 2023, DOI: 10.54941/ahfe1002928, URL: https://doi.org/10.54941/ahfe1002928.

[Bl19] Blumreiter, M.; Greenyer, J.; Garcia, F. J. C.; Klös, V.; Schwammberger, M.; Sommer, C.; Vogelsang, A.; Wortmann, A.: Towards Self-Explainable Cyber-Physical Systems. In (et al., L. B., ed.): 22nd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion, MODELS Companion. IEEE, pp. 543–548, 2019, DOI: 10.1109/MODELS-C.2019.00084.

[Da07] Dautenhahn, K.: Socially intelligent robots: dimensions of human–robot interaction. Philosophical Transactions of the Royal Society B: Biological Sciences 362 (1480), pp. 679–704, 2007, DOI: 10.1098/rstb.2006.2004, URL: https://doi.org/10.1098/rstb.2006.2004.

[FBH23] Fränzle, M.; Bairy, A.; Hajnorouzi, M.: Computational Cognitive Models Meet Reactive Game Theory and Reactive Synthesis: Cognitively Informed Automated Synthesis of Behavioural Strategies Across the Human-Machine Boundary, Subm. to Int. Journal of Human-Computer Studies, 2023.

[Le18] van Leeuwen, E. J. C.; Cohen, E.; Collier-Baker, E.; Rapold, C. J.; Schäfer, M.; Schünemann, B.; Tennie, C.; Vale, G.; Haun, D. B. M.: The development of human social learning across seven societies. Nature Communications 9 (1), p. 2076, 2018, DOI: 10.1038/s41467-018-04468-2, URL: https://doi.org/10.1038/s41467-018-04468-2.

[PA10] Pandey, A. K.; Alami, R.: A framework towards a socially aware Mobile Robot motion in Human-Centered dynamic environment. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. Pp. 5855–5860, 2010, DOI: 10.1109/IROS.2010.5649688.

[RS23] Rakow, A.; Schwammberger, M.: Brake or Drive: On the Relation Between Morality and Traffic Rules when Driving Autonomously. In: Software Engineering 2023 Workshops. Gesellschaft für Informatik e.V., Bonn, pp. 104–115, 2023, DOI: 10.18420/se2023-ws-12.

[Sc18] Schwammberger, M.: An abstract model for proving safety of autonomous urban traffic. Theoretical Computing Science 744, pp. 143–169, 2018, DOI: 10.1016/j.tcs.2018.05.028.

[Sc25] Schwammberger, M.: A Roadmap towards Dynamic Conflict Management for Autonomous Traffic Agents, to be published in Proc. of 36th IEEE Intelligent Vehicles Symposium (IV'25), 2025.

[Sced] Schwammberger, M.; Rakow, A.; Putze, L.; Bairy, A.: Explain it for Safety: Explanations for Risk Mitigation. In (Rauh, A.; Finkbeiner, B.; Kröger, P., eds.): Design and Verification of Cyber-Physical Systems: From Theory to Applications. accepted 2025, to be published.

[Wi15] Wickens, C. D.: Noticing events in the visual workplace: The SEEV and NSEEV models. In: The Cambridge Handbook of Applied Perception Research. Cambridge Handbooks in Psychology, Cambridge University Press, pp. 749–768, 2015, DOI: 10.1017/CBO9780511973017.046.