Responsibility Ascription in Trustworthy Autonomous Systems

Vahid Yazdanpanah[1], Enrico H. Gerding[1], Sebastian Stein[1], Mehdi Dastani[2], Catholijn M. Jonker[3], Timothy J. Norman[1], and Sarvapali D. Ramchurn[1]

[1] University of Southampton, United Kingdom
{v.yazdanpanah,e.gerding,s.stein,t.j.norman,sdr1}@soton.ac.uk
[2] Utrecht University, The Netherlands
m.m.dastani@uu.nl
[3] Delft University of Technology, The Netherlands
c.m.jonker@tudelft.nl

Abstract. To develop and effectively deploy Trustworthy Autonomous Systems (TAS), we face various social, technological, legal, and ethical challenges in which different notions of responsibility can play a key role. In this work, we elaborate on these challenges, discuss research gaps, and show how the multidimensional notion of responsibility can play a key role to bridge them. We argue that TAS requires operational tools to represent and reason about responsibilities of humans as well as AI agents.

Keywords: Trustworthy Autonomous Systems· Multiagent Responsibility Reasoning· Reliable AI· Dynamics of Human-Agent Collectives.

1 Introduction

To develop and effectively deploy Trustworthy Autonomous Systems (TAS) [22, 10], it is crucial to coordinate their behaviour [24], ensure their compatibility with our human-centred social values [27, 19], and design verifiably safe and reliable human-agent collectives [16]. To that end, we face various social, technological, legal, and ethical challenges for which socio-technically expressive notions of responsibility, blameworthiness, accountability, and liability need to be developed. This requires focusing on various interdisciplinary topics as it relates to:

- *Philosophy of AI and Applied Ethics:* Studying the conceptual links between the notion of autonomy and responsibility in human-agent collectives;
- *Sociological Aspects of Agency and Autonomy:* Capturing the social implications of the introduction of autonomous systems into society and conceptualising how different levels of autonomy relate to different notions of responsibility;
- *Legal Studies and Automated Judicial Reasoning Tools:* Formalising legal principles, based on the jurisprudential perspective on responsibility, to govern autonomous systems towards preserving social values and contextual norms; and
- *Multiagent Technologies and Formal Methods:* Developing automated responsibility reasoning tools and decision support services for human-centred autonomous systems.

The need for ensuring trustworthiness of autonomous systems is known and well-argued in the literature [31, 19, 9]. However, as long as we remain at an abstract level and merely discuss how TASs ought to behave (i.e., without clear instructions on potential ways to ensure trustworthiness) the gap will not be bridged. We argue that to ensure TAS, the community requires intermediary notions and operational tools to represent and reason about different facets of trustworthiness in those systems. We require a notion that is, on one hand, rich-enough to capture the aforementioned aspects of TAS and, on the other, computationally implementable. To address this gap, we deem that the multidimensional notion of responsibility in its various forms (e.g., blameworthiness, accountability, sanctionability, and liability) can be used, tailored, and extended for this purpose.

In principle, responsibility necessitates autonomy as this is defined only for an agent with a level of autonomy [2, 6]. From the other side, autonomy is about the capacity of an entity to manifest its agency via performing actions, either communicative or physical [28], and causing change in the environment to reach its desires/preferences [25, 4, 3, 12]. Then agent A causing change and reaching outcome O in the environment indicates *"A's responsibility for O"* [13, 14, 7, 15]. We see that an agent's responsibility can be formulated in terms of the post-conditions of their actions as an *ex post* notion. In other words, in terms of how the execution of affordable strategies resulted in an outcome for which agents are responsible. As a complementary approach, in multiagent settings, the line of research on strategic responsibility and action-state semantics [30, 5, 29, 26] focuses on the strategic capacities of agents or groups of agents with respect to potential situations in prospect. In this view, agents' responsibility is formulated in terms of pre-conditions as an *ex ante* notion. These two forms of retrospective and prospective responsibility are key for conceptualising what van de Poel [23] calls backward- and forward-looking notions of responsibility.

In this proposal, we show how different dimensions and notions of responsibility relate to and can address challenges in design, development, and deployment of trustworthy autonomous systems. This work is an attempt to articulate TAS challenges to which responsibility reasoning and multiagent tools for reasoning about various forms of responsibility can contribute and is a starting point for establishing a research agenda on *"Responsibility Ascription in Trustworthy Autonomous Systems"*.

2 Prospective Responsibility in TAS

In principle, prospective responsibility reasoning is focused on eventualities as potential situations that may be materialised in future and analyses how individual agents or human-agent collectives can or ought to affect such state of affairs in future (e.g., in the sense that, for an upcoming picnic, we say Alice is responsible for driving and Bob is for preparing food). In autonomous systems, prospective responsibility reasoning is crucial, e.g., to ascribe the responsibility for ensuring the safety of an autonomous vehicle system to a capable agent or agent group. This calls for considering the strategic abilities of humans as well as artificial entities in responsibility reasoning and, in turn, in assigning tasks to human-agent collectives. Moreover, responsibility reasoning can be of use to design verifiably reliable autonomous human-agent organisations. Below, we

present TAS challenges that call for novel responsibility reasoning research and discuss desirable requirements to be met.

Challenge 1 *The need for practical and provably sound degrees of responsibility to ensure system reliability and fault tolerance in the technical software development context.*

In real-life autonomous systems, reliability of the system and its ability to handle potential failures are key for social acceptance. The society will not accept the integration of autonomous vehicles unless they show the capacity to perform reliably and in a fault-tolerant manner. One should never expect that all the components in an autonomous system behave as expected, and so one has to put in place overarching methods to ensure reliability. For this, we can rely on formally verifiable responsibility reasoning methods [30, 20]. Following Chockler and Halpern [7], we deem that the notion of responsibility can be a base for conceptualising resilience.

Challenge 2 *The need for operational accountability ascription and task coordination methods in TAS's organisational context.*

In human-agent collectives, where human and artificial agents collaborate towards ensuring goals, it is crucial to put in place mechanisms for balancing the two decision-making types in what Jennings et al. call *flexible autonomy* [16]. In essence, flexible autonomous systems allow "agents to sometimes take actions in a completely autonomous way without reference to humans [type 1], while at other times being guided by much closer human involvement [type 2]". Then the main problem is to understand who is, and to what extent they are, accountable for the outcome of such decisions.

3 Retrospective Responsibility in TAS

In the generic sense of the term, responsibility is mostly understood as a backward-looking, retrospective notion. For instance, imagine a multiagent system with three autonomous vehicles, two pedestrians, and one human-driven vehicle. After the occurrence of a crash (as an already materialised event), retrospective responsibility is to reason about individuals or groups of agents capable of avoiding the crash (in terms of avoidance power [5]) or those who caused it (in terms of causal affirmative power [7]). Computational retrospective responsibility tools can be of use for automated liability determination in TAS, for addressing the so-called responsibility gaps/voids (where a group is determined to be responsible collectively but individuals' share is not clear), and for building sanctioning tools and value-aligned coordination mechanisms to ensure the functionality of human-agent collectives in TAS.

Challenge 3 *The need for verfiably effective tools to address responsibility voids in human-agent collectives and measures to fairly distribute collective-level responsibilities into quantitative individual-level degrees of responsibility.*

Imagine a scenario (adapted from [18]) where a traveller's water canteen is poisoned by one and then emptied by another fellow traveller. The traveller dies of thirst in the middle of the desert. It is clear that the two fellow travellers are responsible as a collective but the extent and the degree of responsibility of each fellow is not clear. This is a stranded case of the so called "*responsibility void*" [1] where linking collective to individual responsibility is a challenge. In the literature on responsibility reasoning, there exist suggestions to adopt cost allocation techniques to ascribe responsibility among agents with respect to their contribution to the collective [30, 11]. While such approaches lead to desirable fairness properties they are not scalable due to their expensive computational complexity.

Challenge 4 *The need for context-aware blameworthiness and accountability reasoning tools as a basis for effective liability measures and to ensure the legality of TAS.*

By giving more autonomy to artificial systems, one cannot see them as object-like tools that merely follow instructions. For instance, a driver-less AV is not receiving direct instructions thus, when collisions occur, a judge cannot simply apply "Qui facit per alium, facit per se" (who acts through another does the act himself) [8, 17, 21] to see the owner as the only responsible agent. It is reasonable that any involved agent with a degree of autonomy takes a degree of blameworthiness and accountability. However, on a basic level, most of our behaviour-enforcement methods are founded on physical regimentation techniques, e.g., to imprison or impose some form of physical restriction, that are neither effective on, nor meaningful for non-human agents. Thus, we deem that, for effective deployment of autonomous systems, it is neither effective nor efficient to rely on non-automated resource-consuming judiciary processes. By doing so, we are automating transportation and manufacturing but need to add much more capacities (human labour, time, and judiciary expertise) to judge each and every incident of failure. This is not an attempt for full automation of the judiciary system but, in contrast, a proposal to capture the capacities of non-human agents, integrate it with social values, and develop human-centred legal decision support tools for TAS.

References

1. Braham, M., van Hees, M.: Responsibility voids. The Philosophical Quarterly **61**(242), 6–15 (2011)
2. Braham, M., van Hees, M.: An anatomy of moral responsibility. Mind **121**(483), 601–634 (2012)
3. Bratman, M.E.: Structures of agency: Essays. Oxford University Press, Oxford, United Kingdom (2007)
4. Bratman, M.E., Israel, D.J., Pollack, M.E.: Plans and resource-bounded practical reasoning. Computational intelligence **4**(3), 349–355 (1988)
5. Bulling, N., Dastani, M.: Coalitional responsibility in strategic settings. In: Computational Logic in Multi-Agent Systems - 14th International Workshop, CLIMA XIV, Corunna, Spain, September 16-18, 2013. Proceedings. pp. 172–189. Springer, Berlin, Heidelberg (2013)
6. Champlin, T.S.: Responsibility. Philosophy **69**(268), 254–255 (1994)
7. Chockler, H., Halpern, J.Y.: Responsibility and blame: A structural-model approach. Journal of Artificial Intelligence Research **22**, 93–115 (2004)

8. Conard, A.: What's wrong with agency. J. Legal Educ. **1**, 540 (1948)
9. Dignum, V.: Responsible Artificial Intelligence - How to Develop and Use AI in a Responsible Way. Artificial Intelligence: Foundations, Theory, and Algorithms, Springer, Switzerland (2019)
10. European Commission: the High-Level Expert Group on AI: Ethics guidelines for trustworthy ai. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (2019), accessed: 2020-09-20
11. Friedenberg, M., Halpern, J.Y.: Blameworthiness in multi-agent settings. In: The Thirty-Third AAAI Conference on Artificial Intelligence. pp. 525–532. AAAI Press, New York (2019)
12. Georgeff, M., Pell, B., Pollack, M., Tambe, M., Wooldridge, M.: The belief-desire-intention model of agency. In: International workshop on agent theories, architectures, and languages. pp. 1–10. Springer, Berlin, Heidelberg (1998)
13. Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach. part i: Causes. The British journal for the philosophy of science **56**(4), 843–887 (2005)
14. Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach. part ii: Explanations. The British journal for the philosophy of science **56**(4), 889–911 (2005)
15. Hamblin, C.: Imperatives, basil black well (1987)
16. Jennings, N.R., Moreau, L., Nicholson, D., Ramchurn, S., Roberts, S., Rodden, T., Rogers, A.: Human-agent collectives. Communications of the ACM **57**(12), 80–88 (2014)
17. Lin, P., Abney, K., Bekey, G.A.: Robot ethics: the ethical and social implications of robotics. Intelligent Robotics and Autonomous Agents series, (2012)
18. McLaughlin, J.A.: Proximate cause. Harvard law review **39**(2), 149–199 (1925)
19. Murukannaiah, P.K., Ajmeri, N., Jonker, C.M., Singh, M.P.: New foundations of ethical multiagent systems. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems. pp. 1706–1710 (2020)
20. Naumov, P., Tao, J.: An epistemic logic of blameworthiness. Artif. Intell. **283**, 103269 (2020)
21. Norman, T.J., Reed, C.: A logic of delegation. Artificial Intelligence **174**(1), 51–71 (2010)
22. Office for Artificial Intelligence: Ethics guidelines for trustworthy AI. https://www.gov.uk/government/publications/a-guide-to-using-artificial-intelligence-in-the-public-sector (2020), accessed: 2020-09-20
23. van de Poel, I.: The relation between forward-looking and backward-looking responsibility. In: Moral responsibility, pp. 37–52. Springer, Dordrecht, The Netherlands (2011)
24. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.F., Breazeal, C., Crandall, J.W., Christakis, N.A., Couzin, I.D., Jackson, M.O., et al.: Machine behaviour. Nature **568**(7753), 477–486 (2019)
25. Rao, A.S., Wooldridge, M.: Foundations of rational agency. In: Foundations of rational agency, pp. 1–10. Springer, Dordrecht, The Netherlands (1999)
26. Reed, C., Norman, T.J.: A formal characterisation of hamblin's action-state semantics. J. Philos. Log. **36**(4), 415–448 (2007)
27. Russell, S.: Human compatible: Artificial intelligence and the problem of control. Penguin, London, UK (2019)
28. Searle, J.R.: How performatives work. Linguistics and philosophy **12**(5), 535–558 (1989)
29. Yazdanpanah, V., Dastani, M.: Distant group responsibility in multi-agent systems. In: International Conference on Principles and Practice of Multi-Agent Systems. pp. 261–278. Springer, Switzerland (2016)
30. Yazdanpanah, V., Dastani, M., Jamroga, W., Alechina, N., Logan, B.: Strategic responsibility under imperfect information. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. p. 592–600 (2019)
31. Zhang, J., Bentahar, J., Falcone, R., Norman, T.J., Şensoy, M.: Introduction to the special section on trust and ai. ACM Trans. Internet Technol. **19**(4) (Nov 2019)