Forward-Looking and Backward-Looking Responsibility Attribution in Multi-Agent Sequential Decision Making

Doctoral Consortium

Stelios Triantafyllou
Max Planck Institute for Software Systems
Saarbrücken, Germany
strianta@mpi-sws.org

ABSTRACT

As AI systems gain more and more agency in modern-day society, the problem of responsibility attribution in AI is no longer just a philosophically interesting one, but a practical one as well. The rise of AI agency means that an increasing number of everyday tasks are now being handled by AI agents. As a result, addressing conceptual and technical challenges of attributing responsibility for the failure of a multi-agent AI system has become urgent. Such challenges are particularly prominent when the temporal dimension of decision making is taken into account. In general, the concept of responsibility attribution may have different meanings depending on the context. In particular, in my research I consider the distinction between forward-looking and backward-looking responsibility. Forward-looking responsibility looks at the future and holds agents accountable for what is expected to happen. On the other hand, backward-looking responsibility looks at the past and holds agents accountable for a specific realization of the system and an outcome of interest. This paper summarizes my contributions on forward- and backward-looking responsibility attribution in multi-agent sequential decision making and describes my future research plans.

KEYWORDS

Responsibility Attribution; Multi-Agent Systems; Markov Decision Processes; Actual Causality; Cooperative Game Theory

ACM Reference Format:

Stelios Triantafyllou. 2023. Forward-Looking and Backward-Looking Responsibility Attribution in Multi-Agent Sequential Decision Making: Doctoral Consortium. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Consider any multi-agent AI system where an agent's decision can be influenced by past decisions of other agents. When such a system fails, an important question arises: to what extent is each agent responsible for this failure? Answering this question could be useful for a number of reasons. For example, we may want to (a) hold agents accountable for the failure of the system [8, 15], (b) understand and explain why the failure happened, and then (c) determine how to avoid such failures in the future [1, 29].

A real-world example of such a multi-agent AI system is autonomous traffic light control (ATLC) [3, 9]. Typically, in ATLC each agent controls one road intersection, and after observing some local information decides on how to schedule the intersection's traffic lights. A common failure mode of an ATLC system occurs when a driver's waiting time, in one of the intersections, exceeds some pre-selected "acceptable" threshold. In such scenarios, however, it is difficult to identify which agent(s) caused the failure, because of the temporal dependencies between the agents' decisions. Did the agent responsible for the specific intersection made a mistake? Or was it impossible for that agent to avoid this situation due to earlier mistakes made by other agents? Answers to these questions can be given by a responsibility attribution method, i.e., a method which assigns a degree of responsibility to each agent that reflects its contributions to the undesirable outcome of interest. These answers can then be utilized by the system designer in efforts to avoid similar failures in the future. For example, when limited resources are available, attention could be focused on modifying the behaviours of agents with higher degrees of responsibility.

Responsibility can be viewed from two perspectives, **forward-looking** and **backward-looking** [27]. The former considers all possible realizations of a system and assigns responsibility to an agent in expectation of what might happen in the future. Going back to the ATLC example, one could ascribe forward responsibility to an agent for the expected total extra time that the drivers will have to wait. In contrast, the backward-looking perspective assigns responsibility to an agent for some specific realization of the system, e.g., for a specific traffic instance. Both notions have been studied before in moral philosophy, law and AI [4, 5, 11, 14, 18, 23].

In my research, I focus on forward-looking and backward-looking responsibility attribution in Multi-Agent Markov Decision Processes with full (MMDPs) or partial (Dec-POMDPs) observability [7, 20]. MMDPs and Dec-POMDPs are two general and widely used frameworks for multi-agent sequential decision making, but for which responsibility attribution had not been studied before. This research direction poses numerous interesting challenges of both a conceptual and technical nature. Conceptual challenges stem from the need to develop responsibility attribution methods that satisfy desirable properties and also align well with human intuition. Furthermore, prior work on responsibility and blame in AI [13, 17] has recognized a number of factors that influence responsibility attribution, including knowledge, intent and others. Incorporating all these factors into a single practical responsibility attribution method in the sequential setting is a non-trivial task. From a technical point of view, there are many interesting challenges related, for example, to uncertainty considerations and the computational complexity of the responsibility attribution problem.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

With the help of my collaborators, I have characterized and tackled various conceptual and technical challenges of responsibility attribution. Looking forward, I aspire to keep progressing in the same direction and apply my findings to a real-world domain.

2 FORWARD-LOOKING RESPONSIBILITY

In my first research paper [25], we consider the task of attributing forward-looking responsibility in cooperative sequential decision making. [1] The formal setting we focus on is Multi-Agent Markov Decision Processes (MMDPs). We also choose to assign responsibility to the agents for the expected discounted return of their joint policy. In other words, we measure an agent's responsibility based on its contributions to the total inefficiency of the system.

We formalize properties and methods for responsibility attribution in the setting of interest, by first considering concepts derived from or inspired by the cooperative game theory literature [6, 28]. Next, we expand the set of desirable properties by including two novel properties that we deem important for responsibility attribution, namely *performance monotonicity* and *Blackstone consistency*.

We show that some of the well-known responsibility attribution methods, such as Shapley Value [22], are not performance monotonic. Roughly speaking, this means that an agent might receive an increased degree of responsibility for adopting a policy that would improve the current inefficiency of the system. To address this issue and guarantee that an agent is always incentivized to reduce the system's inefficiency, we introduce a novel responsibility attribution method that trade-offs explanatory power (by attributing less responsibility to the agents) for performance monotonicity.

Blackstone consistency states that an agent should not receive a higher responsibility just because the agents' policies are not exactly known to the responsibility attribution procedure.[2] To ensure that no agent gets unfairly over-blamed under such uncertainty, we provide algorithms for making all the studied responsibility attribution methods Blackstone consistent.

3 BACKWARD-LOOKING RESPONSIBILITY

In my second piece of work [24, 26], we consider the task of attributing backward-looking responsibility for a specific outcome of interest. Our starting point is a standard approach for attributing responsibility based on **actual causality** [16]:

(1) Pinpoint actual causes, i.e., agents' decisions that were pivotal for the outcome of interest,
(2) Assign a degree of responsibility to each agent based on the found actual causes.

Furthermore, in order to enable causal reasoning, this approach utilizes the Structural Causal Model (SCM) framework [21].

For our formal setting, we establish a connection between SCMs and Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs). This connection allows us to study actual causality and (causal) responsibility attribution in Dec-POMDPs. Under this framework we look at both conceptual and technical sides of the backward-looking responsibility attribution problem.

On the conceptual side [26], we begin by making the observation that existing definitions of actual causality, such as the Halpern and Pearl definition, do not explicitly account for temporal dependencies between agents' decisions. Through examples inspired by moral philosophy and extensive simulation-based experiments we show that this can lead to counter-intuitive actual causes and at the same time negatively affect the responsibility attribution procedure. To address this issue, we introduce a novel definition for actual cause that captures our intuition and prevents such counter-intuitive results. The key characteristic of our definition is that it utilizes a structural component of Dec-POMDPs which models how each agent's decision depends on the agent's interaction history. Other contributions include a family of responsibility attribution methods that extend the well-known Chockler and Halpern approach [10].

Our technical contributions [24] are motivated by the fact that the problem of pinpointing actual causes, and consequently determining the exact responsibility assignments, has shown to be computationally intractable [2, 12]. Thus, in order to apply responsibility attribution in large-scale domains, we would have to find a way to overcome this complexity. To fill this gap, we introduce an efficient search algorithm for approximating the agents' degrees of responsibility under a computational budget. Our algorithm is a variation of the Monte Carlo Tree Search method tailored to the problem of responsibility attribution. It has a number of novel components, such as a new search tree and an elaborate pruning procedure. Note also that our method is generic and can be technically applied to any practical setting modeled as a finite and discrete Dec-POMDP. Finally, we evaluate the efficacy of our approach on a simulation-based test-bed, which consists of three card games.

4 FUTURE WORK

In the future, I plan to focus on backward-looking responsibility attribution in Dec-POMDPs. One project I am interested in is conducting a user-study about the human perception of actual causality and responsibility attribution. In this study, I would like to test how well different actual cause definitions and responsibility attribution mechanisms align with human intuition. Through interactive use cases, I hope to (a) validate my intuition from prior work and (b) gain additional insights that I could utilize when developing future definitions and methods.

Another project that I strongly believe would benefit the field, as it could potentially attract more researchers, is extending my current experimental test-bed. This endeavor would entail creating additional environments suitable for testing different properties and evaluating the efficiency of search methods. By the end of this project I aim to have introduced the first experimental suite for actual causality and responsibility attribution, which would be accessible by researchers even outside of the AI community.

Finally, I am also particularly interested in working on applying responsibility attribution in a real-world domain. Apart from the computational complexity, there are other challenges that need to be addressed first in order to make this goal feasible. Some of these challenges are related to the simplifying assumptions that we made in our previous work. For example, we restricted the underlying model to be finite and discrete. Moreover, for our experiments we assumed a specific class of SCMs, the Gumbel-Max SCMs [19]. Lifting these assumptions is critical for making our work widely applicable in practice.

[1] In the paper, the term *blame* is used instead of responsibility.
[2] The policies might be estimated from data.

ACKNOWLEDGEMENTS

This research was, in part, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360.

REFERENCES

[1] Natasha Alechina, Joseph Y. Halpern, and Brian Logan. 2017. Causality, responsibility and blame in team plans. In *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems*. 1091–1099.

[2] Gadi Aleksandrowicz, Hana Chockler, Joseph Y. Halpern, and Alexander Ivrii. 2017. The computational complexity of structure-based causality. *Journal of Artificial Intelligence Research* 58 (2017), 431–451.

[3] Itamar Arel, Cong Liu, Tom Urbanik, and Airton G. Kohls. 2010. Reinforcement learning-based multi-agent system for network traffic signal control. *IET Intelligent Transport Systems* 4, 2 (2010), 128–135.

[4] Peter M. Asaro. 2007. Robots and responsibility from a legal perspective. *Proc. IEEE* 4, 14 (2007), 20–24.

[5] Christel Baier, Florian Funke, and Rupak Majumdar. 2021. A game-theoretic account of responsibility allocation. *arXiv preprint arXiv:2105.09129* (2021).

[6] Maria-Florina Balcan, Ariel D. Procaccia, and Yair Zick. 2015. Learning cooperative games. In *Proceedings of the 24th International Conference on Artificial Intelligence*. 475–481.

[7] Craig Boutilier. 1996. Planning, learning and coordination in multiagent decision processes. In *TARK*, Vol. 96. Citeseer, 195–210.

[8] Mark Bovens. 2007. Analysing and assessing accountability: A conceptual framework. *European Law Journal* 13, 4 (2007), 447–468.

[9] Chacha Chen, Hua Wei, Nan Xu, Guanjie Zheng, Ming Yang, Yuanhao Xiong, Kai Xu, and Zhenhui Li. 2020. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 3414–3421.

[10] Hana Chockler and Joseph Y. Halpern. 2004. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research* (2004).

[11] Mark Coeckelbergh. 2020. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics* 26, 4 (2020), 2051–2068.

[12] Thomas Eiter and Thomas Lukasiewicz. 2002. Complexity results for structure-based causality. *Artificial Intelligence* 142, 1 (2002), 53–89.

[13] Matija Franklin, Hal Ashton, Edmond Awad, and David Lagnado. 2022. Causal framework of artificial autonomous agent responsibility. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 276–284.

[14] Meir Friedenberg and Joseph Y. Halpern. 2019. Blameworthiness in multi-agent settings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 525–532.

[15] Davide Grossi, Lamber Royakkers, and Frank Dignum. 2007. Organizational structure and responsibility. *Artificial Intelligence and Law* 15, 3 (2007), 223–249.

[16] Joseph Y. Halpern. 2016. *Actual causality*. MiT Press.

[17] Joseph Y. Halpern and Max Kleiman-Weiner. 2018. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[18] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2021. Human perceptions on moral responsibility of AI: A case study in AI-assisted bail decision-making. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[19] Michael Oberst and David Sontag. 2019. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*. 4881–4890.

[20] Frans A Oliehoek and Christopher Amato. 2016. *A concise introduction to decentralized POMDPs*. Springer.

[21] Judea Pearl. 2009. *Causality*. Cambridge University Press.

[22] Lloyd S. Shapley. 1953. A value for n-person games. *Annals of Mathematics Studies* 28 (1953), 307–317.

[23] Steve Torrance. 2008. Ethics and consciousness in artificial agents. *Ai & Society* 22, 4 (2008), 495–521.

[24] Stelios Triantafyllou and Goran Radanovic. 2023. Towards computationally efficient responsibility attribution in decentralized partially observable MDPs. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems AAMAS 2023 (to appear)*.

[25] Stelios Triantafyllou, Adish Singla, and Goran Radanovic. 2021. On blame attribution for accountable multi-agent sequential decision making. *Advances in Neural Information Processing Systems* 34 (2021), 15774–15786.

[26] Stelios Triantafyllou, Adish Singla, and Goran Radanovic. 2022. Actual causality and responsibility attribution in decentralized partially observable Markov decision processes. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 739–752.

[27] Ibo Van de Poel. 2011. The relation between forward-looking and backward-looking responsibility. In *Moral responsibility*. Springer, 37–52.

[28] John Von Neumann and Oskar Morgenstern. 2007. *Theory of games and economic behavior (commemorative edition)*. Princeton University Press.

[29] Vahid Yazdanpanah, Mehdi Dastani, Natasha Alechina, Brian Logan, and Wojciech Jamroga. 2019. Strategic responsibility under imperfect information. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems AAMAS 2019*. IFAAMAS, 592–600.