On Quantitative Analysis of Responsibility in Multiagent Systems

Extended Abstract

Chunyan Mu
University of Aberdeen
Aberdeen, United Kingdom
Chunyan.Mu@abdn.ac.uk

Nir Oren
University of Aberdeen
Aberdeen, United Kingdom
N.Oren@abdn.ac.uk

ABSTRACT

The analysis of responsibility in multi-agent systems is typically binary with an agent either being, or not being responsible for some outcome. In this paper we describe a framework built on a variant of probabilistic alternating-time temporal logic through which we can capture different types of causal responsibility, and provide different measures for the amount of causal responsibility one can ascribe on an agent.

KEYWORDS

MASs, Responsibility Measurement, Formal Verification

ACM Reference Format:
Chunyan Mu and Nir Oren. 2026. On Quantitative Analysis of Responsibility in Multiagent Systems: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/PVFH7189

1 INTRODUCTION

Existing approaches to responsibility analysis [1, 4–6, 9, 11, 14] typically provide a binary assessment as to whether an agent is, or is not, responsible for some outcome. Such analysis is useful in ensuring that the MAS operates effectively, responsibly, and ethically. However, responsibility is often not binary and an agent can be partially responsible for an outcome. A more fine-grained, quantitative understanding of responsibility is important to facilitate features such as blame apportionment. Such a quantitative analysis of responsibility, within the context of coalition plans, is the focus of this paper.

Our aim is to provide a more precise understanding of overall responsibility for outcomes over time, aligning with the notions proposed in [12]. We consider a coalition of agents acting within a MAS, and ascribe different types of responsibility as follows.

- An agent has *Causal Active Responsibility* for some outcome when other agents cannot act to prevent that outcome from occurring, but the responsible agent could have acted differently to prevent the outcome from happening.
- An agent has *Causal Passive Responsibility* for some outcome if, when other agents' actions are kept fixed, the agent acting differently would have prevented the outcome from occurring.
- An agent has *Causal Contributive Responsibility* if it is part of a group of agents whose actions together led to an outcome, while the coalition could not have achieved that outcome without the agent's actions.

In the remainder of this paper we describe our underlying model and extensions to probabilistic ATL which allow us to capture the three types of responsibility. We then detail three metrics which measure how much causal (active/passive/contributive) responsibility an agent bears for a given outcome in a joint plan. and show that these yield very different results. Due to space constraints, we only provide a high level description in this extended abstract. Full details of our model and results can be found at [10].

2 THE MODEL

We model our approach as a transition system. From some state, agents select an action. All agent actions are then combined as a single *joint* action which governs which new state the system transitions to. In our formalism, (joint) actions have probabilistic outcomes, i.e., a probability distribution describes the resultant state following the execution of a joint action. *Histories* describe a possible evolution of a system from some starting state, and are sequences of states and actions. Clearly, given the probability distribution of actions, we can compute the likelihood of a history occurring.

Agents follow (memoryless) strategies, which allow them to probabilistically select an action given a state. Thus, given an initial state and a strategy profile, we may obtain a set of histories which describe the possible evolutions of a system. A joint plan for a subset of agents identifies those histories compatible with those agents' strategies. We can then say that two such plans are equivalent (for some coalition of agents, and for some state) if agents follow the same actions when executing the plans.

We can model the system using a variant of probabilistic ATL [2, 8]. More specifically we consider finite (i.e., bounded) paths, yielding the following syntax.

$$\begin{aligned} \phi &::= a \mid \neg\phi \mid \phi \wedge \phi \mid \langle\Gamma\rangle[\psi] \mid \langle\Gamma\rangle\mathbf{P}_{\bowtie p}[\psi] \\ \psi &::= \bigcirc\phi \mid \phi\mathbf{U}_{\leq k}\phi \end{aligned}$$

Here a is an atomic proposition and Γ is a set of agents. Thus, $\langle\Gamma\rangle[\psi]$ denotes that a set of agents has a joint strategy to enforce ψ, while $\langle\Gamma\rangle\mathbf{P}_{\bowtie p}[\psi]$ states that a set of agents has a strategy to ensure that the formula ψ is satisfied with a likelihood governed by $\bowtie p$ (where $\bowtie$ is an inequality symbol). Finally, note the bounded until formula which enforces the usual until condition up to states of depth k. Using this until operator allows one to derive bounded eventually ($\diamond_{\leq k}$) and always ($\Box_{\leq k}$) operators.

We model the three types of responsibility as first class elements in our logic, by adding the following to the semantics:

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/PVFH7189

$\langle\Gamma\rangle\mathrm{CAR}_{i,\pi}(\psi)$, $\langle\Gamma\rangle\mathrm{CPR}_{i,\pi}(\psi)$ and $\langle\Gamma\rangle\mathrm{CCR}_{i,\pi}(\psi)$. These respectively denote that with respect to a coalition Γ and joint plan π, agent i has CAR/CPR/CCR with respect to outcome ψ. We can easily define the semantics for these operators; for example, CAR is formalised as follows.

$s \models_{\mathcal{G}} \langle\Gamma\rangle\mathrm{CAR}_{i,\pi}(\psi)$ iff

$\forall\pi' \in \mathrm{Plan}_{\pi}^{\langle\{i\}\rangle}(s).\forall\boldsymbol{\sigma}'_{\Gamma} \in \pi'(\Gamma).\forall\rho \in \mathrm{Hist}_{\mathcal{G}}^{\boldsymbol{\sigma}'_{\Gamma}}(s).\rho \models_{\mathcal{G}} \psi \wedge$

$\exists\pi'' \in \mathrm{Plan}_{\pi}^{\langle\mathrm{Ag}\rangle}(s).\forall\boldsymbol{\sigma}''_{\Gamma} \in \pi''(\Gamma).\forall\rho \in \mathrm{Hist}_{\mathcal{G}}^{\boldsymbol{\sigma}''_{\Gamma}}(s).\rho \not\models_{\mathcal{G}} \psi$

While it is easy to define algorithms which check whether an agent has each type of responsibility, the complexity of undertaking this check is in PSPACE.

3 MEASURING CAUSAL RESPONSIBILITY

Our measures for causal responsibility revolve around the number of states, and/or their likelihoods, where some property is or is not satisfied. In the current work, we consider three measures.

The *proportional measure* counts the number of histories where some property is satisfied, and normalises this by the total number of histories.

The *probabilistic measure* we sum the probabilities of each history where the property is satisfied, and again normalise by the total probability of all histories. The probabilistic measure is more nuanced than the proportional measure. However, consider a situation where an agent must ensure that some state holds once every 100 iterations, versus a situation where the agent must ensure that the state holds every iteration, and the agent is able to achieve the outcome with a likelihood of 50% for any iteration. In the latter case, the probability of success as the system runs towards infinity is $\lim_{t\to\infty} \frac{1}{2}^t = 0$, while in the former case it is $\lim_{t\to\infty} \frac{1-1}{2^100}^t = 0$. Thus, while it is (intuitively) easier to maintain the desired state once every hundred iterations, the probabilistic measure cannot discriminate between these two cases.

To overcome this we introduce an *entropy measure*. The entropy of a language (of finite words) $\mathcal{L}$ is defined as $\mathcal{H}(\mathcal{L}) = \limsup_{n\to\infty} \frac{\log_2(1+|\mathcal{L}^n|)}{n}$. Returning to the previous example, we find that the entropy measure returns 0.5 for the case where the state needs to be maintained every iteration, and $\frac{1.99}{2}$ when it needs to be visited at least every hundredth iteration, reflecting that the former is harder to achieve than the latter.

To measure an agent i causal active responsibility we denote by $\mathcal{L}^{+}_{\mathrm{CAR}_i}$ those histories where an outcome is satisfied when an agent acts alone, and by $\mathcal{L}^{-}_{\mathrm{CAR}_i}$ those histories where the coalition of agents avoids the outcome. The proportional measure is obtained by computing $C = \frac{\mathcal{L}^{+}_{\mathrm{CAR}_i}}{\mathcal{L}^{+}_{\mathrm{CAR}_i}+\mathcal{L}^{-}_{\mathrm{CAR}_i}}$, while the probabilistic measure computes the value by summing the total likelihood of each positive history. If the longest history has length k, then the entropy measure returns $\frac{\log_2 C}{k}$. Similar calculations can be used to derive an agent's passive and contributive responsibility, we refer the reader to [10] for further details.

We evaluated our approach on a simple variant of the repeated prisoner's dilemma. Here, agents can cooperate or defect. If both agents cooperate, they are rewarded (denoted by the state `reward`), if both defect, they receive a fine (`fine`), and if only i cooperates, the cooperator receives a payoff ($\texttt{payoff}_\texttt{i}$). Each agent defects with a likelihood of 0.25 and cooperates with a likelihood of 0.75.

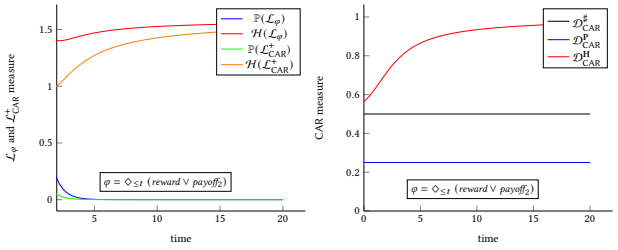

Figure 1: CAR **for the measures over time**

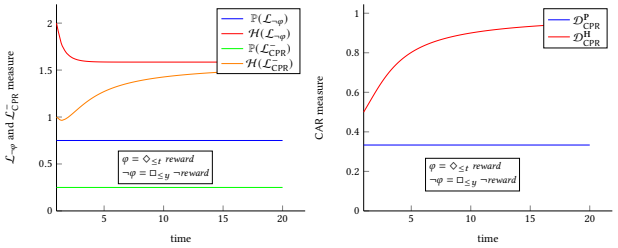

Figure 2: CPR **over time**

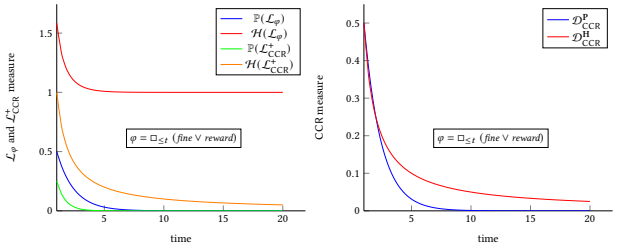

Figure 3: CCR **over time.**

Figures 1-3 show the results of the measures across CAR, CPR and CCR for a variety of formulae over time (e.g., $\mathcal{D}^{\mathbb{P}}_{\mathrm{CCR}}$ denotes the collective causal responsibility for an agent under the probabilistic measure. We observe that the measures return different results, demonstrating their importance.

4 CONCLUSIONS

We introduced measures for causal responsibility which incorporate the nature of an outcome (achievable/unachievable or avoidable/unavoidable), offering context-aware evaluation of an agent's active, passive or contributive responsibility. The measures are sensitive to variations in agent behaviours, and demonstrate differences in responsibility based on different joint plans and outcomes over time. The measures align with different intuitive expectations of responsibility, and are useful across diverse MAS scenarios.

As future work we will investigate the link between our measures and Shapely values [13], potentially through the use of strategic logic [3, 7]. We also intend to examine tradeoffs between responsibility and coalition performance, as well as system resilience.

REFERENCES

[1] N. Alechina, J. Y. Halpern, and B. Logan. 2017. Causality, Responsibility and Blame in Team Plans. In *AAMAS*. ACM, 1091–1099.

[2] R. Alur, T. A. Henzinger, and O. Kupferman. 2002. Alternating-time temporal logic. *J. ACM* 49, 5 (2002), 672–713.

[3] B. Aminof, M. Kwiatkowska, B. Maubert, A. Murano, and S. Rubin. 2019. Probabilistic Strategy Logic. In *IJCAI*. 32–38.

[4] C. Baier, F. Funke, and R. Majumdar. 2021. A Game-Theoretic Account of Responsibility Allocation. In *IJCAI*. ijcai.org, 1773–1779.

[5] C. Baier, F. Funke, and R. Majumdar. 2021. Responsibility Attribution in Parameterized Markovian Models. In *AAAI*. AAAI Press, 11734–11743.

[6] C. Baier, R. van den Bossche, S. Klüppelholz, J. Lehmann, and J. Piribauer. 2024. Backward Responsibility in Transition Systems Using General Power Indices. In *AAAI*. AAAI Press, 20320–20327.

[7] K. Chatterjee, T. A. Henzinger, and N. Piterman. 2010. Strategy logic. *Inf. Comput.* 208, 6 (2010), 677–693.

[8] T. Chen and J. Lu. 2007. Probabilistic Alternating-time Temporal Logic and Model Checking Algorithm. In *FSKD*. IEEE Computer Society, 35–39.

[9] M. Gladyshev, N. Alechina, M. Dastani, and D. Doder. 2023. Group Responsibility for Exceeding Risk Threshold. In *KR*. 322–332.

[10] Chunyan Mu and Nir Oren. 2024. Measuring Responsibility in Multi-Agent Systems. arXiv:2411.00887 [cs.MA] https://arxiv.org/abs/2411.00887

[11] P. Naumov and J. Tao. 2021. Two Forms of Responsibility in Strategic Games. In *IJCAI*. ijcai.org, 1989–1995.

[12] T. Parker, U. Grandi, and E. Lorini. 2023. Anticipating Responsibility in Multiagent Planning. In *ECAI*, Vol. 372. 1859–1866.

[13] Lloyd S Shapley. 1953. A Value for n-Person Games. In *Contributions to the Theory of Games II*, Harold W. Kuhn and Albert W. Tucker (Eds.). Princeton University Press, 307–317.

[14] V. Yazdanpanah, M. Dastani, W. Jamroga, N. Alechina, and B. Logan. 2019. Strategic Responsibility Under Imperfect Information. In *AAMAS*. International Foundation for Autonomous Agents and Multiagent Systems, 592–600.