Causality in Multi-Agent Systems

Doctoral Consortium

Sylvia S. Kerkhove
Utrecht University
Utrecht, Netherlands
s.s.kerkhove@uu.nl

ABSTRACT

Over the last couple of years, causality has become of bigger interest to the AI community. It has, among other things, been used to generate explanations of black-box models. Despite this interest, research into causality in strategic multi-agent systems settings has been lacking. This project intends to develop methods to study causality in multi-agent systems, with the goal of determining accountability of system outcomes. In order to do this, we first discuss what we understand by causality. We then introduce a first attempt at developing a causal model for a strategic multi-agent setting. Finally, we discuss how causal questions could be answered more efficiently using abstraction techniques.

KEYWORDS

Causality, Multi-Agent Systems, Strategic Behaviour

ACM Reference Format:
Sylvia S. Kerkhove. 2025. Causality in Multi-Agent Systems: Doctoral Consortium. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Consider a self-driving car that, in order to be permitted on the road, still requires the driver to pay attention. Now imagine that this car crashes into a tree that fell over the road during a storm. Should we blame the driver for this? After all, they must have been distracted when they should have been paying attention. Or should we maybe blame the car? It should in normal circumstances have been capable of driving safely on its own, the human driver having to pay attention was just a safety measure. Or can we blame the storm for felling the tree? After all, there are not normally trees on the road. In order to determine the answer to these questions, we must first ask ourself what *caused* the collision.

Causality has been studied since antiquity [8], but the modern view of causality dates back to Hume's work in the 18th century [10]. Nowadays, research distinguishes two different notions of causality, *type causality* and *actual causality* [8]. Type causality focuses on general statements and typically tries to use causal notions to predict future events. We find type causality in statements like: "being distracted while driving causes accidents," and "a bad posture will cause problems with your spine." Actual causality on the other hand focuses on specific events and generally tries to explain why an event happened. We find actual causality in statements like: "the driver hitting the gas pedal instead of the brakes caused the collision," and "my neck hurts because I was in a car-crash."

There have been quite a few different attempts at giving a formal definition for actual causality over the years (see for example [2, 3, 6, 7, 9, 12]). Most formal definition use Pearl's structural causal model framework [14]. In such causal models the world is described in terms of variables, which are divided into a set of *exogenous* and a set of *endogenous variables*. The exogenous variables are variables whose values are determined by causes outside of the model [8]. These causes are not really of interest to the modeller and often, the values of the exogenous variables will be determined through a probability distribution [8]. The causes of endogenous variables are of interest to the modeller. The endogenous variable values are consequently determined by either exogenous variables, or other endogenous variables [8]. Every endogenous variable has a corresponding structural equation that specifies how the value of this variable is determined by the other variables [14].

People have claimed that a formal definition of actual causality requires two main components [15]. Firstly, we should require that the potential cause and the event have both actually happened. Secondly, there has to be some condition that says that if the cause had happened differently, then the event should also have been different, this is a type of *counterfactual* argument. This is reflected in most definitions (e.g. [2, 6, 9]) and there is indeed experimental evidence that people reason about causality in such counterfactual ways [5]. The simplest definition of causality, the but-for definition, just simply takes these two requirements [8]. In our earlier example this would lead to us concluding that the driver being distracted is a cause of the collision, just in case that there was a collision and the driver was distracted, *and* that if the driver had not been distracted there also would not have been a collision. However, the but-for definition is seen as too restrictive by most people, but the different formal definitions of actual causality are testament to how hard it is to find a formal definition that people agree on. The most well known definition of actual causality is the Halpern-Pearl Definition, but it has seen several iterations over the years [8].

Definitions of actual causality have been used in AI to provide explanations of automated decisions and to determine which part of a system should be held responsible for unwanted outcomes. For example, to provide explanations of the decisions of a deep neural network, its structure could be described as a structural causal model [13]. The user could use this to reason about how each component of the neural network affects the outcome.

Because agents can causally influence each other in multi-agent settings, it is difficult to assign responsibility for certain outcomes to a specific group of agents. This project aims to develop techniques

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

to make these causal influences more explicit. For this, we will introduce a way to combine strategic multi-agent settings with causality. Afterwards, we will study how to make this combination more efficient, using abstraction techniques for causal models.

2 DEFINING RESPONSIBILITY IN MULTI-AGENT SYSTEMS USING CAUSALITY

In a multi-agent system it is important to be able to determine where the responsibility for an outcome lies. If we do not know what part of the system was responsible for an unwanted outcome, we can also not know how to avoid such an outcome in the future.

While responsibility has been defined using just agent strategies [16], other approaches argue that an agent has to have been a cause of the outcome in order for it to be responsible for it [4]. Conversely, in the agent strategy approach, a group of agents is seen as responsible for an outcome if they had a strategy to prevent it. We have introduced a causal concurrent game structure (causal CGS) in order to attempt to unify these two approaches [11]. A concurrent game structure (CGS) is a type of transition system. It consists of states and possible actions for agents in each state. CGS allow us to reason about agent strategies in multi-agent settings.

To construct a causal concurrent game structure, we start with a recursive causal model (meaning that the variables do not have cyclic dependencies) where the variables can attain finitely many values and where these values are deterministically determined. In this model, we partition the endogenous variables into a set of agent and a set of environment variables. Agent variables are those variables that will be directly influenced by the agents of the model, and the environment variables are all other variables. We use the ancestral relations of the variables in the acyclic causal graph of the model to determine the order in which the agents are allowed to take their actions in the causal CGS. We see interventions on the causal model as possible agent actions.

We can show that a group of agents is a but-for cause of a certain outcome in a structural causal model if and only if they had a strategy to prevent it in the causal CGS based on the SCM [11]. We have also shown a more complicated result for a group of agents that is a cause according to the HP definition. These results show a relation between responsibility as defined by having an ability to avoid an outcome and as defined by being a cause of the outcome.

So far, when developing the causal CGS, we have only looked at deterministic and recursive models. However, in many practical use cases, the causal models are probabilistic, so we would like to see whether we can extend our model to a probabilistic setting as well. We believe that this is not necessarily straightforward. For one, concurrent game structures are also deterministic, so we need to either use a different type of model (Markov games for example), or extend CGS to deal with probabilities. A second problem would be that many probabilistic causal models have real variable values. Our causal CGS only works for finite-valued variables and in order to extend our approach to infinite values we might want to think about grouping values together in order to give us finitely many options again.

Another possible extension would be allowing for non-recursive models, as assuming that the model is acyclic reduces our expressiveness. Assuming that all relations are acyclic makes it impossible for us to model mutual dependencies among agents. Nevertheless, this extension would not be simple, as determining the order in which agents get to take actions would be harder. Moreover, evaluating the states would be problematic as variable values will depend on each other. Still, we would like to see if adding a temporal component to the model could help mitigate these problems. Therefore, while we feel that it would be natural to attempt to extend our definition of causal CGS to probabilistic and non-recursive models as well, we foresee several challenges when defining these extensions.

3 EFFICIENTLY REPRESENTING CAUSALITY IN MULTI-AGENT SYSTEMS

Structural causal models can become quite complex when there are a lot of variables involved. Moreover, our causal CGS is polynomial in size of the original causal model. It would hence be beneficial to develop ways to abstract the causal model, so that it becomes less complex and more easily interpretable.

There has been recent research into abstracting causal models by Beckers and Halpern [1]. In their paper they define a higher-level model to be an abstraction of a lower-level model if there exists a surjective function, subject to some extra constraints, between the variable values of the lower-level model and the variable values of the higher-level model.

While this definition seems promising, there are no formal results to support that this definition is a good one. This is in general something that seems to be lacking in papers that aim to equate causal models. We hence intend to develop methods that will make it more clear what we mean when we say that a certain causal abstraction technique is a good one. We intend to do this by defining how causal relations should carry over between models.

However, comparing causal relations in different models is not a simple task. The models have different variables, so while ideally one would like that if a variable was a cause of an event in the first model, it is also a cause of this event in the second model, this is impossible, given that the variable (or even the event) may not exist in the second model. Nevertheless, in the Beckers & Halpern paper, the variable in the original model is related to variables in the second model through the surjective function on the variable values [1]. We aim to see whether we can use this function to see whether the variables it relates have similar causal relations.

4 CONCLUSION

We have discussed why having a decent framework to discuss causality in multi-agent systems is needed. A first framework has been introduced, using concurrent game structures, but this can only be used for deterministic, recursive models. We intend to extend this framework in the future and to develop techniques to help with the evaluation of large-scale models.

ACKNOWLEDGMENTS

This publication is part of the CAUSES project (KIVI.2019.004) of the research programme Responsible Use of Artificial Intelligence, which is financed by the Dutch Research Council (NWO) and ProRail.

REFERENCES

[1] Sander Beckers and Joseph Y Halpern. 2019. Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, Vol. 33. 2678–2685.

[2] Sander Beckers and Joost Vennekens. 2018. A principled approach to defining actual causation. *Synthese* 195, 2 (Feb. 2018), 835–862. https://doi.org/10.1007/s11229-016-1247-1

[3] Alexander Bochman. 2018. Actual Causality in a Logical Setting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 1730–1736. https://doi.org/10.24963/ijcai.2018/239

[4] Hana Chockler and Joseph Y Halpern. 2004. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research* 22 (2004), 93–115.

[5] Tobias Gerstenberg, Noah D. Goodman, David A. Lagnado, and Joshua B. Tenenbaum. 2015. How, whether, why: Causal judgments as counterfactual contrasts.. In *CogSci*. https://cicl.stanford.edu/papers/gerstenberg2015how.pdf

[6] Maksim Gladyshev, Natasha Alechina, Mehdi Dastani, Dragan Doder, and Brian Logan. 2023. Dynamic Causality. In *ECAI 2023-26th European Conference on Artificial Intelligence, including 12th Conference on Prestigious Applications of Intelligent Systems, PAIS 2023-Proceedings*. IOS Press BV.

[7] N. Hall. 2007. Structural equations and causation. *Philosophical Studies* 132, 1 (Jan. 2007), 109–136. https://doi.org/10.1007/s11098-006-9057-9

[8] Joseph Y Halpern. 2016. *Actual causality*. MIT Press.

[9] Joseph Y. Halpern and Judea Pearl. 2005. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science* 56, 4 (2005), 843–887. https://doi.org/10.1093/bjps/axi147

[10] D Hume. 1958. An enquiry concerning human understanding (1748) Reprinted by Open Court Press. *LaSalle, IL* (1958).

[11] Sylvia S. Kerkhove, Mehdi Dastani, and Natasha Alechina. 2025. Causes and Strategies in Multiagent Systems. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May19-23*. IFAAMAS, upcoming.

[12] Emiliano Lorini. 2023. A Rule-Based Modal View of Causal Reasoning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, Edith Elkind (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3286–3295. https://doi.org/10.24963/ijcai.2023/366 Main Track.

[13] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. 2020. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter* 22, 1 (2020), 18–33.

[14] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.

[15] Brad Weslake. 2015. A Partial Theory of Actual Causation. *British Journal for the Philosophy of Science* (2015).

[16] 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.