Eliciting Causal Knowledge from Agents

Matteo Ceriscioli

School of Electrical Engineering and Computer Science (EECS)
Oregon State University, Corvallis (OR), USA
ceriscim@oregonstate.edu

Abstract

Causal discovery is the task of learning a causal model from a source of information. Traditionally, the community has focused on algorithms that infer causal models from observational and/or interventional data, while alternative approaches have been only marginally explored. The proposed work aims to contribute to the theoretical foundations connecting agent-based systems with causal modeling, and to identify conditions under which newly developed causal discovery algorithms can be applied to elicit causal knowledge from agents.

Introduction

Over the past thirty years, the field of causality has advanced substantially, providing rigorous frameworks to represent and reason about causality and clarifying the types of questions causal knowledge enables us to answer (Pearl 2009). On top of these developments, a variety of techniques that rely on causal models have been proposed, making the ability to reliably construct causal models increasingly important. While both theoretical results and algorithms for learning causal models from observational data (Spirtes 2001) and interventional data (Li, Jaber, and Bareinboim 2023) have been well established, the use of alternative sources of information for constructing causal models remains largely underexplored. At the same time, rapid advances in training procedures and the development of robust AI systems have shown that these systems can achieve strong generalization performance. Recent work shows that a necessary condition for such robustness is that AI systems acquire causal knowledge of their environment (Richens and Everitt 2024; Ceriscioli and Mohan 2025a). As these systems are increasingly deployed and become more robust, the development of algorithms that both elicit their learned causal knowledge and provide theoretical guarantees comparable to those of traditional causal discovery methods becomes an increasingly promising direction.

The goal of my thesis is to uncover the connections between an agent's behavior and its causal understanding of the environment in which it operates. This perspective enables contributions to the field of causal discovery in a significantly orthogonal direction to prior work: eliciting causal knowledge from agents.

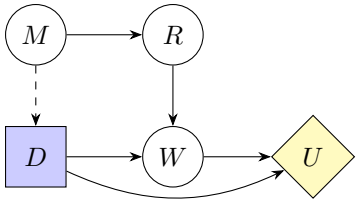

Figure 1: A CID representing a mediated decision task. Nodes represent variables, full arrows represent cause-effect relationships, and dashed arrows indicate what an agent observes before making a decision. In this example, an AI agent controls a sprinkler activation corresponding to the decision D. M represents the month of the year, R rain, and W the humidity of the lawn. The utility node U is associated with a utility function that rewards the agent if the lawn is humid and penalizes water wastage. Adapted from (Ceriscioli and Mohan 2025b).

Adaptable Agents Learn Causal Knowledge

The first part of the thesis consists in building the mathematical framework and provide solid justifications for why and how it is possible to extract causal knowledge from agents. In this context, an agent is any entity that maps percepts to actions, and maximizes expected utility (Russell and Norvig 1995). Most of popular AI systems satisfy this very general definition, for example image classification systems, smart sprinkler controllers, or self-driving cars. Causal Influence Diagrams (CIDs) (Everitt et al. 2021), a causal version of Influence Diagrams (Howard and Matheson 1984), are one way to simultaneously represent these kind of decision tasks and causal relationships. An example of a CID representing a decision task involving a smart sprinkler controller is provided in Figure 1.

Prior research (Richens and Everitt 2024) shows that if an agent is capable of adapting to distribution shifts, it must have learned the causal structure of its environment in single-agent unmediated tasks, i.e., in tasks represented with CIDs where the decision cannot influence the utility through variables of the environment. Specifically, the no-mediation assumption translates into the CID structure as requiring that

the only directed path from the decision node (blue) to the utility node (yellow) is the edge $D \to U$. While this no-mediation assumption significantly restricts the applicability of these results to a very limited set of decision tasks, since many real-world AI applications involve tasks where mediation exists, they also provide initial evidence that there exists a link between an agent ability to adapt to shifts in the environment and its causal understanding of the environment.

Recognizing the potential of this approach, and aiming to build a broader theoretical foundation for extracting causal knowledge from agents, we first extended the framework to the more general family of mediated tasks and outlined preliminary applications in multi-agent systems (Ceriscioli and Mohan 2025c). We then examined sequential decision tasks (Ceriscioli and Mohan 2025b), and ultimately refined these directions into a comprehensive theoretical framework in Ceriscioli and Mohan (2025a).

Moreover, acknowledging that causal understanding is necessary to adapt to changes in the environment opens up new questions. For instance, how do different training paradigms (e.g., traditional reinforcement learning vs. continual learning) or specific techniques influence the acquisition of causal knowledge while simultaneously contributing to the learning of robust policies?

Causal Discovery with Agents

Once the theoretical groundwork and justification have been established, the natural next step is to identify settings in which scalable and theoretically grounded algorithms can be developed to elicit causal knowledge from agents. Such settings can be characterized along several dimensions: the family of decision tasks (mediated vs. unmediated, one-shot vs. sequential), the way the agent and its capabilities are represented (e.g., optimal policy oracles vs. bounded-performance policy oracles), the choice of causal modeling framework for the decision task, and the assumptions about prior knowledge, including causal knowledge, available to the user performing discovery. As in traditional causal discovery, and as observed in Ceriscioli and Mohan (2025a), different configurations of these elements lead to different identifiability results for the underlying causal model. In the literature, substantial effort has gone into matching the right class of causal models to the right class of discovery problems, for example, Partial Ancestral Graphs (PAGs) (Spirtes 2001) for settings with latent confounding and selection bias. Similarly, progress on developing new algorithms for causal discovery with agents will require a deeper understanding of the conditions under which causal knowledge can be extracted, and in particular, of which combinations of models and assumptions yield the strongest results.

Inspecting Agents' Causal Belief

A third aspect of extracting causal knowledge from agents arises when an agent is suboptimal, or when some components of the environment are only partly relevant to its task. In such cases, it may be valuable to learn the agent's causal beliefs about its environment and the factors shaping its decision-making. For example, one might ask which components of the environment an agent considers safe to manipulate in order to achieve its goals, with potential applications for limiting power-seeking behaviors and improving the safety of AI systems. Additionally, understanding an agent's causal beliefs could inform its training by guiding exploration of the state space toward areas that maximize information gain about the environment's causal structure.

Conclusions

This thesis project advances the idea that agents can serve as novel sources of causal knowledge. It aims to establish a theoretical foundation connecting adaptability to causal understanding, contributing to a systematic justification for when and why causal knowledge can be extracted from agents. It then outlines the requirements for scalable algorithms that elicit both causal structure and agents' causal beliefs. Together, these contributions chart a new direction for causal discovery, where agents themselves become active participants in revealing the causal structures of their environments, with implications for both the theory of causality and the design of safer, more robust AI systems.

References

Ceriscioli, M.; and Mohan, K. 2025a. Agents Robust to Distribution Shifts Learn Causal World Models Even Under Mediation. In *Advances in Neural Information Processing Systems*.

Ceriscioli, M.; and Mohan, K. 2025b. Causal Discovery via Adaptive Agents in Multi-Agent and Sequential Decision Tasks. In *The Seventeenth Workshop on Adaptive and Learning Agents*.

Ceriscioli, M.; and Mohan, K. 2025c. Causal Discovery with Adaptable AI Agents. In *AAAI 2025 Workshop on Artificial Intelligence with Causal Techniques*.

Everitt, T.; Carey, R.; Langlois, E. D.; Ortega, P. A.; and Legg, S. 2021. Agent Incentives: A Causal Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Howard, R. A.; and Matheson, J. E. 1984. Influence Diagrams. *Readings on the Principles and Applications of Decision Analysis*.

Li, A.; Jaber, A.; and Bareinboim, E. 2023. Causal discovery from observational and interventional data across multiple environments. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. USA: Cambridge University Press, 2nd edition.

Richens, J.; and Everitt, T. 2024. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*.

Russell, S.; and Norvig, P. 1995. *Artificial Intelligence: A Modern Approach*. Series in Artificial Intelligence. Englewood Cliffs, NJ: Prentice-Hall.

Spirtes, P. 2001. An Anytime Algorithm for Causal Inference. In Richardson, T. S.; and Jaakkola, T. S., eds., *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*. PMLR.