Causal Explanations for Sequential Decision Making Under Uncertainty

Extended Abstract

Samer B. Nashed
University of Massachusetts Amherst, USA
snashed@cs.umass.edu

Saaduddin Mahmud
University of Massachusetts Amherst, USA
smahmud@cs.umass.edu

Claudia V. Goldman
General Motors, Herzliya Pituach, Israel
claudia.goldman@gm.com

Shlomo Zilberstein
University of Massachusetts Amherst, USA
shlomo@cs.umass.edu

KEYWORDS

reasoning under uncertainty; causal reasoning; explainable AI

ACM Reference Format:
Samer B. Nashed, Saaduddin Mahmud, Claudia V. Goldman, and Shlomo Zilberstein. 2023. Causal Explanations for Sequential Decision Making Under Uncertainty: Extended Abstract. 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

As autonomous decision making becomes ubiquitous, researchers agree that developing trust is required for adoption and proficient use of AI systems [20, 35, 39], and it is widely accepted that autonomous agents that can explain their decisions help promote trust [4, 10, 28]. However, there are many challenges in generating such explanations. Consider, for example, an autonomous vehicle (AV) stopped behind a truck. Passengers may wonder whether the AV is waiting for the truck to move, waiting for an opportunity to pass the truck, or dealing with some technical problem. Generating suitable explanations of such a system is hard due to the complexity of planning, which may involve large state spaces, stochastic actions, imperfect observations, and complicated objectives. Furthermore, useful explanations must somehow reduce the internal reasoning process to a form understandable by a non-expert user.

We introduce a novel framework (see also [31]) for causal explanations of stochastic, sequential decision-making systems built on the well-studied structural causal model (SCM) paradigm for causal reasoning [9]. Our unified framework can identify multiple, semantically distinct explanations of agent actions — something not previously possible. We establish exact methods and several approximation techniques for causal inference on Markov decision processes using this framework, followed by results on the applicability of the exact methods and some run-time bounds. We discuss scenarios that illustrate the framework's flexibility and experiments with human subjects that confirm the benefits of this approach.

We operate on several established conclusions of philosophers, psychologists, and cognitive scientists as to the purpose and logical mechanisms that underpin explanations [29, 30] — chiefly that requests for explanations are often motivated by a mismatch between the mental model of the requester and a logical conclusion based on observation [11, 12, 14, 15, 22, 37], and that explanations often require counterfactual analysis [13, 21, 24, 25], which in turn requires causal determination [23, 34, 38]. There are several computational paradigms for causal analysis, including those based on conditional logic [8, 19], and statistics [7]. Among the most well-studied paradigms is the structural causal model [9].

Research on explanations of stochastic planners, such as MDPs, is relatively sparse, but there are several notable existing efforts. Elizalde et al. [6] identify important state factors by looking at how the value function would change were they to perturb that state factor's value, and Khan et al. [17] present a technique to explain policies for factored MDPs by analyzing the expected occupancy frequency of states with extreme reward values. Similarly, Juozapaitis et al. [16] analyze how extreme reward values impact action selection in decomposed-reward RL agents. Wang et al. [36] try to explain policies of partially observable MDPs by communicating the relative likelihoods of different events or levels of belief.

Our framework, based on SCMs, applies causal analysis to sequential decision-making agents by creating an SCM representing the computation needed to derive a policy for a Markov decision process (MDP) and applying causal inference to identify variables that cause certain agent behavior, which can then be used to generate explanations. This framework provides two main benefits. First, it is theoretically sound, based on concepts and formalisms from the causality literature, while most existing approaches use heuristics. Second, it is flexible, allowing us to identify multiple types of explanans, whereas existing approaches often explain events in terms of a single set of variables in the decision-making model. For example, they may use *only* state factors or *only* reward variables, whereas we may use any set. Furthermore, we offer several approximate techniques for large problems or problems where the topology of the causal graph prevents exact inference. We conclude with results from a user study comparing the proposed method to existing, heuristic methods and we find statistically significant preferences in favor of explanations generated via causal reasoning.

2 STRUCTURAL CAUSAL MODELS FOR MDPS

We construct a causal model of the computation that solves for the policy of an MDP and then use this model to determine causes for agent actions. In the general case, this process follows four steps: (1) A causal graph is generated from the relevant MDP components, (2) The resulting graph is converted into a layered causal graph, (3) The layered graph is pruned to remove any irrelevant nodes and

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.

Method	F	R	T	V	Causal?
Elizalde et al. (2009)	Yes	-	-	-	No
Russell and Santos (2019)	Yes	-	-	-	No
Khan et al. (2009)	-	Yes	-	-	No
Juozapaitis et al. (2019)	-	Yes	-	-	No
Betram et al. (2018)	-	Yes	-	-	No
Wang et al. (2016)	-	-	Yes	-	No
Madumal et al. (2020)	Yes	Yes	-	-	Yes
Proposed	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Table 1: Comparison of method applicability

edges given the parameters of the causal query, and (4) A recursive algorithm identifies sets of causal variables in the graph. This provides a principled, general framework for causal inference on MDPs while simultaneously supporting several types of explanations.

Although it is possible to create a single, monolithic causal graph that simultaneously represents all components of the MDP tuple, this is not helpful since it does not afford any additional types of inference and is much less computationally efficient. Here, we give an example of an SCM constructed by only analyzing the state-factors of an MDP. In this work, we use the following definition of cause from Halpern and Pearl [9].

Definition 1. *Let $X \subseteq V$ be a subset of the endogenous variables, and let x be a specific assignment of values for those variables. Given an event ϕ, defined as a logical expression, for instance $\phi = (\neg a \wedge b)$, a* weak cause *of ϕ satisfies the following conditions:*

(1) *Given the context $U = u$ and $X = x$, ϕ holds.*
(2) *Some $W \subseteq (V \setminus X)$ and some $\bar{x}$ and w exist such that:*
A) using these values produces $\neg\phi$.
B) for all $W' \subseteq W$, $Z \subseteq V \setminus (X \cup W)$, where $w' = w|W'$ and $z = Z$ given $U = u$, ϕ holds when $X = x$.

Here, item 2 B) is saying that, given context $U = u$, $X = x$ alone is sufficient to cause ϕ, independent of some other variables W. If π_{sa} is a variable that is true when action a may be taken in state s, then in an MDP, a natural choice for ϕ is a subset of the variables π_{sa}. For example, if action a is taken in state s instead of a', we have

$$\phi = \langle [\pi(s) = a], [\pi(s) = a'] \rangle = \langle \textsc{True}, \textsc{False} \rangle.$$

However, it is less clear how to define potential explanans, denoted by X. Intuitively, we often define X as being, for example, the set of all state factors, the set of all reward variables, or the set of all values for states h actions away. That is, we tend to define X according to some semantic type. The following example model can answer queries about the causality of state factors. Here, we let $\mathcal{U} = \emptyset$, and

$$\mathcal{V} = \pi_{sa} \cup s \cup f_i \qquad \forall s \in S, \forall a \in A, \forall i \in \{1, \ldots, n\}$$

where f_i denotes the ith state factor. Finally, $\mathcal{M}$ is composed of the following three sets.

$$\mathcal{M}_F := f_i = f_i^t, \qquad \forall i \in \{1, \ldots, n\}.$$

Here f_i^t is the value of state factor i at time t. A given set of state factors $\langle f_1, \ldots, f_n \rangle \in f$ determines the state $s \in S$.

$$\mathcal{M}_S := [s = s_i] = [f_1 = f_1^i] \wedge \ldots \wedge [f_n = f_n^i], \qquad \forall s \in S.$$

Last, we have equations representing action selection.

$$\mathcal{M}_A := [\pi(s) = a] = \pi_{sa} \wedge s \qquad \forall s \in S, a \in A.$$

Thus we define $\mathcal{M} := \mathcal{M}_F \cup \mathcal{M}_S \cup \mathcal{M}_A$. In general, this definition of SCMs for state factors permits exact inference regardless of the underlying MDP topology. Importantly, this causal model represents a *fixed* policy. While this model *cannot* change state factors to produce a different policy, it *can* understand how state factors affect action selection for a given policy.

3 RESULTS

Here we present results from an algorithm similar to that presented by Bertossi et al. (2020), based on the concept of responsibility from Chockler and Halpern (2004). This algorithm iterates directly through possible weak causal sets and then progressively checks larger sets W for assignments w that satisfy Def. 1. In addition to finding weak causal sets consistent with Def. 1, it also provides a ranking over causal sets. The purpose of this study is to show how (1) our approach can handle *semantically different* types of causal queries (see Table 1, not all modeled in this abstract), corresponding to different conceptions of MDP explanation in the literature, and (2) formal definitions of causality identify sensible explanans.

We identify 4 general types of explanation in the literature, each focusing on one component of the MDP tuple: *state factors* (F) [6, 33], *rewards* (R) [2, 16, 17], *transitions* (T) [36], and *future states and values* (V) [32]. These papers define metrics and algorithms particular to their type and lead us to define the following.

Definition 2. *Y-type explanations use explanans $x \subset Y$. For example, F-type explanations use the set of state factors.*

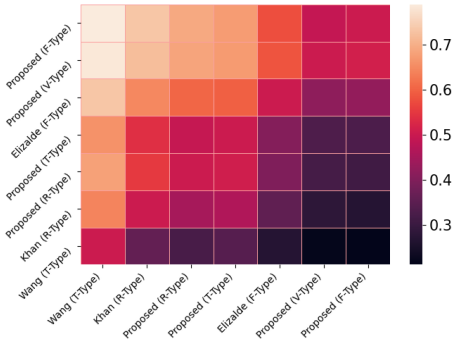

Figure 1: Preference likelihoods of explanation methods. Color indicates the probability that explanations generated using the row-method are preferred to those generated by the column-method.

Here, we describe the results of a study investigating whether users tend to prefer explanations generated using causal reasoning over those generated using heuristics. In total, 189 participants aged 18-65 were shown simulated driving scenarios [18] where a car acts based on a policy from an MDP. After each action participants were shown automatically generated explanations and asked to rank them, producing preference ordering. The explanations included three baselines: [6] (F-type), [17] (R-type), and [36] T-type as well as all four types of explanation generated by our method. Every explanation was presented using the same basic template: "The car <took action> because <explanan 1>, ..., <explanan N>."

Figure 1 summarizes our findings: for every explanation type, users prefer the explanations generated via causal reasoning. We applied the Mann-Whitney U-test [27] to each pair of generation methods (21 in total), using an initial α-value of 0.5, and a Bonferroni-corrected [3] α-value of 0.0024. We detected the following preference ordering with p-values below 0.0001.

1) Prop-F ~ Prop-V ≻ Elizalde ≻ Prop-R ~ Prop-T ≻ Khan ≻ Wang

Here, $A \succ B$ denotes a strict preference for A over B, and ~ denotes preference equality. We believe the overall preference for causal explanations is due to their consistent relevance across all scenarios. Please see the full paper for more discussion and analysis [31].

REFERENCES

[1] Leopoldo Bertossi, Jordan Li, Maximilian Schleich, Dan Suciu, and Zografoula Vagena. 2020. Causality-based Explanation of Classification Outcomes. *arXiv preprint arXiv:2003.06868* (2020).

[2] Josh Bertram and Peng Wei. 2018. Explainable deterministic MDPs. *arXiv preprint arXiv:1806.03492* (2018).

[3] Carlo Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8 (1936), 3–62.

[4] Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright, and Michael Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science* 19, 3 (2018), 259–282.

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

[6] Francisco Elizalde, Enrique Sucar, Julieta Noguez, and Alberto Reyes. 2009. Generating explanations based on Markov decision processes. In *Mexican International Conference on Artificial Intelligence*. Springer, 51–62.

[7] David A Freedman. 2007. Statistical models for causation. *The SAGE Handbook of Social Science Methodology* (2007), 127–146.

[8] Laura Giordano and Camilla Schwind. 2004. Conditional logic of actions and causation. *Artificial intelligence* 157, 1-2 (2004), 239–279.

[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* 52, 3 (2005), 613–622.

[10] Bradley Hayes and Julie A Shah. 2017. Improving robot controller transparency through autonomous policy explanation. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 303–312.

[11] Fritz Heider. 1958. *The psychology of interpersonal relations*. Wiley, New York.

[12] Germund Hesslow. 1988. The problem of causal selection. *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality* (1988), 11–32.

[13] Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107, 1 (1990), 65.

[14] Denis J Hilton. 1996. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning* 2, 4 (1996), 273–308.

[15] Denis J Hilton and Ben R Slugoski. 1986. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review* 93, 1 (1986), 75.

[16] Zoe Juozapaitis, Anurag Koul, Alan Fern, Martin Erwig, and Finale Doshi-Velez. 2019. Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on Explainable Artificial Intelligence*.

[17] Omar Khan, Pascal Poupart, and James Black. 2009. Minimal sufficient explanations for factored Markov decision processes. In *International Conference on Automated Planning and Scheduling (ICAPS)*, Vol. 19.

[18] Edouard Leurent. 2018. An Environment for Autonomous Driving Decision-Making. https://github.com/eleurent/highway-env.

[19] David Lewis. 1974. Causation. *The Journal of Philosophy* 70, 17 (1974), 556–567.

[20] Michael P Linegang, Heather A Stoner, Michael J Patterson, Bobbie D Seppelt, Joshua D Hoffman, Zachariah B Crittendon, and John D Lee. 2006. Human-automation collaboration in dynamic mission planning: A challenge requiring an ecological approach. *Human Factors and Ergonomics Society Annual Meeting* 50, 23 (2006), 2482–2486.

[21] Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements* 27 (1990), 247–266.

[22] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in Cognitive Sciences* 10, 10 (2006), 464–470.

[23] Tania Lombrozo. 2010. Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology* 61, 4 (2010), 303–332.

[24] Tania Lombrozo. 2012. Explanation and abductive inference. In *The Oxford Handbook of Thinking and Reasoning*, K J Holyoak and R G Morrison (Eds.). Oxford University Press, 260–276.

[25] John Leslie Mackie. 1980. *The cement of the universe: A study of causation*. Clarendon Press.

[26] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Explainable reinforcement learning through a causal lens. In *AAAI Conference on Artificial Intelligence*, Vol. 34. 2493–2500.

[27] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.

[28] Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human-agent teaming for Multi-UxV management. *Human Factors* 58, 3 (2016), 401–415.

[29] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[30] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *ACM Conference on Fairness, Accountability, and Transparency*. 279–288.

[31] Samer B. Nashed, Saaduddin Mahmud, Claudia V. Goldman, and Shlomo Zilberstein. 2022. Causal Explanations for Sequential Decision Making Under Uncertainty. *arXiv preprint arXiv:2205.15462* (2022).

[32] Hadrien Pouget, Hana Chockler, Youcheng Sun, and Daniel Kroening. 2020. Ranking Policy Decisions. *arXiv preprint arXiv:2008.13607* (2020).

[33] Jacob Russell and Eugene Santos. 2019. Explaining reward functions in Markov decision processes. In *Thirty-Second International FLAIRS Conference*.

[34] Wesley C Salmon. 2006. *Four decades of scientific explanation*. University of Pittsburgh press.

[35] Kristen Stubbs, Pamela J Hinds, and David Wettergreen. 2007. Autonomy and common ground in human-robot interaction: A field study. *IEEE Intelligent Systems* 22, 2 (2007), 42–50.

[36] Ning Wang, David V Pynadath, and Susan G Hill. 2016. The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 997–1005.

[37] Joseph Jay Williams, Tania Lombrozo, and Bob Rehder. 2013. The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General* 142, 4 (2013), 1006.

[38] James Woodward. 2005. *Making things happen: A theory of causal explanation*. Oxford university press.

[39] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *ACM Conference on Fairness, Accountability, and Transparency*. 295–305.