Causal Explanations for Sequential Decision Making (Abstract Reprint)

Samer B. Nashed[1], Saaduddin Mahmud[1], Claudia V. Goldman[2], Shlomo Zilberstein[1]

[1]University of Massachusetts Amherst, USA
[2]Hebrew University, Israel

Abstract Reprint. This is an abstract reprint of the journal article by Nashed, Mahmud, Goldman, and Zilberstein (2025).

Abstract

Stochastic sequential decision-making systems such as Markov decision processes and their variants are increasingly used in areas such as transportation, healthcare, and communication. However, the ability to explain these systems outputs to non-technical end users has not kept pace with their widespread adoption. This paper addresses that gap by extending prior work and presenting a unified framework for generating causal explanations of agent behavior in sequential decision-making settings, grounded in the structural causal model (SCM) paradigm. Our framework supports the generation of multiple, semantically distinct explanations for agent actions capabilities that were previously unattainable. In addition to introducing a novel taxonomy of explanations for MDPs to guide empirical investigation, we develop both exact and approximate causal inference methods within the SCM framework. We analyze their applicability and derive run-time bounds for each. This leads to the proposed algorithm, MeanRESP, which operates flexibly across a spectrum of approximations tailored to external constraints. We further analyze the sample complexity and error rates of approximate MeanRESP, and provide a detailed comparison of its outputs under varying definitions of responsibility with popular Shapley-value-based methods. Empirically, we performed a series of experiments to evaluate the practicality and effectiveness of the proposed system, focusing on real-world computational demands and the validity and reliability of metrics for comparing approximate and exact causal methods. Finally, we present two user studies that reveal user preferences for certain types of explanations and demonstrate a strong preference for explanations generated by our framework compared to those from other state-of-the-art systems.

References

Nashed, S. B.; Mahmud, S.; Goldman, C. V.; and Zilberstein, S. 2025. Causal Explanations for Sequential Decision Making. *Journal of Artificial Intelligence Research*, 83.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.