Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Journal Track

Explain It as Simple as Possible, but No Simpler – Explanation via Model Simplification for Addressing Inferential Gap (Abstract Reprint)

Sarath Sreedharan [1], **Siddharth Srivastava** [2] and **Subbarao Kambhampati** [2]
[1]Colorado State University, Fort Collins, CO 80524, USA
[2]Arizona State University, Tempe, AZ 85281, USA
ssreedh3@colostate.edu

Abstract Reprint. This is an abstract reprint of a journal article by [Alfano *et al.*, 2024].

Abstract

One of the core challenges of explaining decisions made by modern AI systems is the need to address the potential gap in the inferential capabilities of the system generating the decision and the user trying to make sense of it. This inferential capability gap becomes even more critical when it comes to explaining sequential decisions. While there have been some isolated efforts at developing explanation methods suited for complex decision-making settings, most of these current efforts are limited in scope. In this paper, we introduce a general framework for generating explanations in the presence of inferential capability gaps. A framework that is grounded in the generation of simplified representations of the agent model through the application of a sequence of model simplifying transformations. This framework not only allows us to develop an extremely general explanation generation algorithm, but we see that many of the existing works in this direction could be seen as specific instantiations of our more general method. While the ideas presented in this paper are general enough to be applied to any decision-making framework, we will focus on instantiating the framework in the context of stochastic planning problems. As a part of this instantiation, we will also provide an exhaustive characterization of explanatory queries and an analysis of various classes of applicable transformations. We will evaluate the effectiveness of transformation-based explanations through both synthetic experiments and user studies.

References

[Alfano *et al.*, 2024] Gianvincenzo Alfano, Andrea Cohen, Sebastian Gottifredi, Sergio Greco, Francesco Parisi, and Guillermo R. Simari. Credulous acceptance in high-order argumentation frameworks with necessities: An incremental approach. *Artificial Intelligence*, 333:104159, 2024.