Explainable Agency in Reinforcement Learning Agents

Prashan Madumal
The University of Melbourne
Victoria, Australia
pmathugama@student.unimelb.edu.au

Abstract

This thesis explores how reinforcement learning (RL) agents can provide explanations for their actions and behaviours. As humans, we build *causal models* to encode cause-effect relations of events and use these to explain *why* events happen. Taking inspiration from cognitive psychology and social science literature, I build *causal* explanation models and explanation dialogue models for RL agents. By mimicking human-like explanation models, these agents can provide explanations that are *natural* and *intuitive* to humans.

Introduction

Explainable AI (XAI), a research agenda explored since the era of expert systems (Chandrasekaran, Tanner, and Josephson 1989), has seen renewed interest in recent years with the advent of regulations for Artificial Intelligence (AI) systems. A key pillar of XAI is *explanation*, a justification given for decisions and actions of the system.

However, much research and practice in XAI pays little attention to *people* as the intended users of these systems (Miller 2018). If we are to build systems that are capable of providing 'good' explanations, it is plausible that explanation models should mimic models of human explanation (Madumal 2019).

This sets the premise for my thesis, where I take inspiration from the wealth of pertinent literature in cognitive psychology that explores the nature of explanations and how people understand them. As humans, we view the world through a causal lens (Sloman 2005), building mental models with causal relationships to act in the world, to understand new events and also to *explain* events. Importantly, causal models give people the ability to consider *counterfactuals* — events that did not happen, but could have under different situations. Although this notion of causal explanation is also backed by literature in philosophy and social psychology (Hilton 2007), causality and counterfactuals are only just becoming more prevalent in XAI. Further, compared to the burst of XAI research in supervised learning, explainability in reinforcement learning is hardly explored. Thus, my thesis will explore how causal explanation models can be developed for reinforcement learning agents.

Completed Research

Dialogue Models for Explanation

As noted noted by Hilton (1990), understanding how humans engage in conversational explanation is imperative when building causal explanation models. To this end, I explored how dialogue models can capture the interaction sequences of causal explanations of agents (Madumal et al. 2019). We introduced a dialogue model and an interaction protocol that is grounded on data obtained from different types of explanations in actual conversations.

We derived the model by analysing **398** explanation dialogues using grounded theory across six different dialogue types. We formalised the explanation dialogue model using the *agent dialog framework* (ADF) (McBurney and Parsons 2002), then validated the model in a human-agent experiment with 101 explanation dialogues. The proposed model (Madumal et al. 2019) is general enough to be applied to a wide variety of human-agent interaction domains since it is formalised and presented through a finite state machine. Model was empirically evaluated through a user study in a human-agent setting in a competitive gaming environment, where the agent aids one player by predicting the opponent's strategies and giving explanations of predictions through interacting with the human.

Causal Models for Explanation

In making sense of the world, we build *causal models* in our mind to encode cause-effect relations of events and use these to *explain* why new events happen by referring to counterfactuals — things that did not happen. I use causal models to derive causal explanations of the behaviour of model-free reinforcement learning agents.

I introduce an *action influence* model (Madumal et al. 2020) for model-free reinforcement learning (RL) agents and provide a formalisation of the model using structural causal models (Halpern and Pearl 2005). Action influence models approximate the causal model of the environment relative to actions taken by an agent. Figure 1 shows the action influence graph of a Starcraft II agent. Our approach

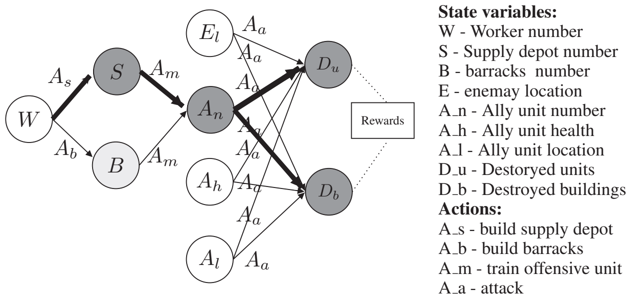

Figure 1: Action influence graph of a Starcraft II agent, causal chain for action A_s is shown in bold.

differs from previous work in explainable RL in that we use causal models to generate *contrastive* explanations for *why* and *why not* questions, which previous models lack. Given assumptions about the direction of causal relationships between variables, during the policy learning process, we also learn the quantitative influences that actions have on variables. Which enable our model to reason approximately about counterfactual states and actions. We define how to generate explanations for 'why?' and 'why not?' questions from the action influence model.

We computationally evaluated our approach on 6 RL benchmarks domains using 6 different RL algorithms. Results indicate that these models are robust and accurate enough to perform task prediction (Hoffman et al. 2018, p.12) with a negligible performance impact. We conducted a human study using the implemented model for RL agents trained to play the real-time strategy game *Starcraft II*. Experiments were run for **120** participants, in which we evaluated the participants' performance in task prediction, explanation satisfaction, and trust. Results show that our model performs better than the tested baseline, but its impact on trust is not statistically significant (De Graaf and Malle 2017; Madumal 2019).

Current Work and Future Directions

Causal Abstraction in Explanation For a large causal graph, explanations generated through an *action influence* model risks overwhelming the explainee. My current work explores how action influence models can be abstracted based on the level of the epistemic knowledge of the explainee, that can provide the necessary granularity. We use *causal ordering* (Iwasaki and Simon 1994) as the theoretical foundation to explore this concept.

Learning Causal Models for Explanation To generate explanations from an action influence model, the causal structure of the domain need to be given beforehand. But this is not always possible, especially in fully autonomous domains. Thus, future work of my thesis involves approximating the causal structure of the domain in an RL agent for explanation.

Conclusion

To make reinforcement learning agents explainable, this thesis follows a causal approach taking inspiration from cognitive science and social psychology. As causal explanations closely resembles human explanations, they have the potential to be natural and intuitive.

References

Chandrasekaran, B.; Tanner, M. C.; and Josephson, J. R. 1989. Explaining control strategies in problem solving. *IEEE Intelligent Systems* (1):9–15.

De Graaf, M. M., and Malle, B. F. 2017. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*.

Halpern, J. Y., and Pearl, J. 2005. Causes and explanations: A structural-model approach. part II: Explanations. *The British journal for the philosophy of science* 56(4):889–911.

Hilton, D. J. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107(1):65.

Hilton, D. 2007. Causal explanation. *Social psychology: Handbook of basic principles* 232–253.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Iwasaki, Y., and Simon, H. A. 1994. Causality and model abstraction. *Artificial intelligence* 67(1):143–194.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2019. A grounded interaction protocol for explainable artificial intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 1033–1041.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2020. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Madumal, P. 2019. Explainable agency in intelligent agents: Doctoral consortium. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2432–2434.

McBurney, P., and Parsons, S. 2002. Games that agents play: A formal framework for dialogues between autonomous agents. *Journal of logic, language and information* 11(3):315–334.

Miller, T. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

Sloman, S. 2005. *Causal models: How people think about the world and its alternatives*. Oxford University Press.