Explainable Agency in Intelligent Agents

Doctoral Consortium

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

ABSTRACT

Explainability of intelligent agents has gained attention in recent years with their widespread utilization in society. Most work in Explainable AI (XAI) pay little attention to the humans that interact with these models, which risks resulting in unsatisfactory explanations. Theories of explainability and the nature of explanation has been widely explored in cognitive psychology and philosophy. This thesis aims to incorporate these insights to build explainable models and interfaces that can provide better and sufficient explanations to the interacting human.

KEYWORDS

Explainable AI; Interpretable Machine Learning ; Causal Explanation

ACM Reference Format:
Prashan Madumal. 2019. Explainable Agency in Intelligent Agents. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Artificial intelligence (AI) systems that aim to be transparent about their decisions must have understandable explanations that clearly justify their decisions. This is especially true in scenarios where people are required to make critical decisions based on the outcomes of an AI system. An appropriate explanation can promote trust in the system, allowing better human-AI cooperation [14]. Explanations also help people to reason about the extent to which, if at all, they should trust the provider of the explanation.

However, much research and practice in explainable AI uses the researchers' intuitions of what constitutes a 'good' explanation rather than basing the approach on a strong understanding of how people define, generate, select, evaluate, and present explanations [10, 11]. Most modern work on Explainable AI, such as in autonomous agents [3, 17] and interpretable machine learning [2], does not discuss the cognitive and the social aspect of the explanations.

As Miller [10, pg 10] notes, the process of Explanation involves two processes: (a) a *Cognitive process*, namely the process of determining an explanation for a given event, called the *explanandum*, in which the causes for the event are identified and a subset of these causes is selected as the explanation (or *explanans*); and (b) the *Social process* of transferring knowledge between explainer and explainee, generally an interaction between a group of people, in which the goal is that the explainee has enough information to understand the causes of the event. The cognitive process and the social process of explanation can be directly mapped to the explainable model and the explanation interface respectively and together forms the research agenda of this project.

Explanation naturally occurs as a continuous interaction, which gives the interacting party the ability to question and interrogate explanations. This allows the explainee to clear doubts about the given explanation by further interrogations and user-driven questions. Further, the explainee can express contrasting views about the explanation that can set the premise for an argumentation based interaction. This type of iterative explanation can provide richer and satisfactory explanations as opposed to one-shot explanations. This forms the motivation for an explanation interface that cater argumentation and continuous interaction.

De Graaf [1] note that humans attribute human traits, such as beliefs, desires, and intentions, to intelligent agents, and it is thus a small step to assume that people will seek to explain agent behaviour using human frameworks of explanation. We hypothesise that AI explanation models with designs that are influenced by human explanation models have the potential to provide more intuitive explanations to humans and therefore be more likely to be understood and accepted. We suggest it is easier for the AI to emulate human explanations rather than expecting humans to adapt to a novel and unfamiliar explanation model. Humans observe and understand the world through a causal lens and represent knowledge in causal relations and models [12, 13], thus having causal models of explanation can intuitively provide human-like explanations. This research will build explainable causal models that augment the underlying AI models.

2 DIALOGICAL EXPLANATION AS AN EXPLANATION INTERFACE

Understanding how humans engage in conversational explanation is a prerequisite to building an explanation model, as noted by Hilton [5]. While there are mature existing models for explanation dialogs [15, 16], these are idealised conceptual models that are not grounded on or validated by data, and seem to lack iterative features like cyclic dialogs. We introduce a dialogue model and an interaction protocol that is based on data obtained from different types of explanations in actual conversations.

We derive our model by analysing 398 explanation dialogues using grounded theory [4] across six different dialogues types. We formalize the explanation dialogue model using the *agent dialogue framework* (ADF) [9], then validate the model in a human-agent study with 101 explanation dialogues. We propose that by following

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

a data-driven approach to formulate and validate, our model more accurately defines the structure and the sequence of an explanation dialog and will support more natural interaction with human audiences than explanations from existing models.

Our model is general enough to be applied to a wide variety of human-agent interaction domains as it is formalized 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. Further evaluations using different domains will be done with an improved model that can cater parallel dialogues, which was found as a weakness through the conducted study.

3 CAUSAL EXPLAINABLE MODELS

Humans understand and act upon the world through a causal lens, in that we observe, do and imagine events in a cause and effect manner [12]. Thus humans expect explanations to questions they ask that adhere to these causal models. According to Pearl [12] these 3 distinct levels of cognitive ability can be depicted using the 'causal ladder'. Lower most level (association) can answer 'what if..' questions and the second level (intervention) can answer 'what if/how..' questions.

However only the topmost layer (counterfactual) can answer 'why' questions such as 'Why P event occurred instead of Q event'. Lim et al. [8] asserts that Why and Why not questions are the most demanded questions asked from intelligible systems, thus any truly explainable system should thrive to obtain the ability to answer why questions. In order to reliably provide explanations to why questions a causal model that can reason counterfactually is required, which most current explainable systems lack.

We propose causal models as a mechanism to augment the underlying AI model, and explore under the context of model-free reinforcement learning. In the reinforcement learning problem, we model the the causal graph as a DAG which constitutes of state variables and rewards as nodes, actions as edges. Moreover we define actions as interventions that is done to the causal graph. We assume causal markov condition to the graph. Causal relations of the variables has to be known prior in the given domain. Then we introduce algorithms to generate explanations for why and why not questions. We leverage the policy of the agent in a given snapshot to obtain the state variable values, and then apply them to the causal graph. We can then generate the explanation by 1) Obtaining the *explanandum* (variable/action that user needs explanation; and 2) Generating the *explanans* (explanation) by traversing the causal graph through to root reward node. We formalise the problem through structural causal equations [12], with the variable relation modeled as a linear relation. For evaluation, we choose Starcraft II (a real-time strategy game) as the domain for its complexity rich causal structure.

Human experiments ($n = 120$) we done to evaluate the model, where the participants were given explanations about the behaviour of a reinforcement learning agent playing Starcraft II. After the explanations are given, participants were presented with several gameplay videos of the agent and were asked to predict behaviour of the agent in the next instance. Through this task prediction [6, pg 12], participants understanding of the agent gained through explanations were measured. We also measured the 'satisfaction' of explanations using the scale described in [6, pg 39] and 'trust' using the scale described in [6, pg 49]. Results indicate that our causal explanation model is statistically significantly better compared to previous models of explanation, indicated by task prediction scores. Although the self reported explanation 'satisfaction' metrics were significantly better for our model, 'trust' metrics did not show any significant differences between models. We attribute this to the limited time participants had to interact with the Stacraft II agent.

4 FUTURE DIRECTIONS

Humans are likely to have different or partially complete causal models of the agent's domain. Causal explanations should change in order to facilitate the model updates of the human by explaining the causal relations. This research will also explore how to generate explanations when the model of the explainee is partially complete and algorithms to update explainee's model.

Explanations naturally have different levels of abstractions, which can be selective depending on the explainee. We will explore abstraction in the context of causal models, where explanation generation will use abstracted causal models. Humans use abstraction in causal models extensively, where a causal template is used to extrapolate into more detailed causal models [13].

One major challenge in causal explanation generation is the lack of an accurate causal model for a given domain, thus we will explore approaches of causal discovery that enable the generation of causal models. Approaches that aim to find a dynamic Bayesian Network in a reinforcement learning problem will be done as an initial step [7].

This research aims to contribute to the growing literature of XAI through a human-centered approach using causal and dialogical explanation models.

5 ACKNOWLEDGEMENT

The author would like to thank A/Prof. Tim Miller (University of Melbourne), Prof. Liz Sonenberg (University of Melbourne) and Prof. Frank Vetere (University of Melbourne) for their guidance and support.

REFERENCES

[1] MM de Graaf and Bertram F Malle. 2017. How People Explain Action (and Autonomous Intelligent Systems Should Too). In *AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction*.

[2] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. Ml (2017), 1–13. arXiv:1702.08608 http://arxiv.org/abs/1702.08608

[3] Maria Fox, Derek Long, and Daniele Magazzeni. 2017. Explainable Planning. *IJCAI - Workshop on Explainable AI* (2017).

[4] Barney G Glaser and Anselm L Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Vol. 1. 271 pages. https://doi.org/10.2307/2575405 arXiv:arXiv:gr-qc/9809069v1

[5] Denis J. Hilton. 1991. A Conversational Model of Causal Explanation. *European Review of Social Psychology* 2 (1991), 51–81. https://doi.org/10.1080/14792779143000024

[6] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *arXiv preprint arXiv:1812.04608* (2018).

[7] Anders Jonsson and Andrew Barto. 2007. Active learning of dynamic bayesian networks in markov decision processes. In *International Symposium on Abstraction, Reformulation, and Approximation*. Springer, 273–284.

[8] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2119–2128.

[9] Peter McBurney and Simon Parsons. 2002. Games That Agents Play: A Formal Framework for Dialogues between Autonomous Agents. *Journal of Logic, Language and Information* 11, 3 (01 Jun 2002), 315–334. https://doi.org/10.1023/A:1015586128739

[10] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (2019), 1–38.

[11] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum; Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. In *IJCAI-17 Workshop on Explainable AI (XAI)*. 36.

[12] Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

[13] Steven Sloman. 2005. *Causal models: How people think about the world and its alternatives*. Oxford University Press.

[14] Serena Villata, Guido Boella, Dov M. Gabbay, and Leendert Van Der Torre. 2013. A socio-cognitive model of trust using argumentation theory. *International Journal of Approximate Reasoning* 54, 4 (2013), 541–559. https://doi.org/10.1016/j.ijar.2012.09.001

[15] Douglas Walton. 2011. A dialogue system specification for explanation. *Synthese* 182, 3 (2011), 349–374. https://doi.org/10.1007/s11229-010-9745-z

[16] Douglas Walton. 2016. *A Dialogue System for Evaluating Explanations*. Springer International Publishing, Cham, 69–116. https://doi.org/10.1007/978-3-319-19626-8_3

[17] Michael Winikoff. 2017. Debugging Agent Programs with Why?: Questions. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS '17)*. IFAAMAS, 251–259.