Explaining Agent Preferences & Behavior: Integrating Reward-Decomposition & Contrastive-Highlights

Extended Abstract

Yael Septon
Technion - I.I,T
Haifa, Israel
yael123@campus.technion.ac.il

Yotam Amitai
Technion - I.I,T
Haifa, Israel
yotama@campus.technion.ac.il

Ofra Amir
Technion - I.I,T
Haifa, Israel
oamir@technion.ac.il

ABSTRACT

Explainable reinforcement learning methods aim to help elucidate agent policies and their underlying decision-making processes. One such method is reward decomposition, which aims to reveal an agent's preferences in a specific world-state by presenting its expected utility decomposed to different components of the reward function. While this approach quantifies the expected decomposed rewards for alternative actions, it does not demonstrate the outcomes of these alternative actions in terms of the behavior of the agent. This work introduces "Contrastive Highlights", a novel local explanation method that visually compares the agent's chosen behavior to an alternative choice of action in a contrastive manner. We conducted user studies comparing participants' understanding of agents' preferences based on either reward decomposition, contrastive highlights, or a combination of both approaches. Our results show that integrating reward decomposition with contrastive highlights significantly improved participants' performance compared to using each of the approaches separately.

KEYWORDS

Explainable AI; Human-AI Interaction; Deep Reinforcement Learning

ACM Reference Format:
Yael Septon, Yotam Amitai, and Ofra Amir. 2023. Explaining Agent Preferences & Behavior: Integrating Reward-Decomposition & Contrastive-Highlights: 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

This work focuses on helping participants develop accurate mental models of an RL agent's preferences and understand the trade-offs between its alternative actions. To this end, we develop "contrastive highlights", a novel local explanation method that draws inspiration from global policy summaries, but focuses on local decisions – which action to take in state s. Similar to policy summaries for single agents, contrastive highlights convey agent behavior by showing trajectories of the agent acting in the environment. However, while policy summaries only show the actions chosen by the agent, contrastive highlights show both the trajectory beginning with the chosen action, as well as a simulated *contrastive* one depicting the agent's trajectory had it chosen a different action in the same world-state and then continued to follow its original policy. This approach aims to provide more information regarding the decision made by the agent by showing the outcomes of the chosen action and an alternative action side-by-side.

Contrastive highlights can be integrated into policy summaries, and can also complement other local explanation methods. In particular, this approach naturally complements reward decomposition which quantifies the agent's expectations regarding the utility of different actions, by demonstrating the outcomes of alternative actions in the environment. Contrastive highlights answer the question "What if?" while reward decomposition answers the question "why?" an action was chosen. In addition, the contrastive nature of the explanation is in line with the literature on explanation in the social sciences which show that people typically provide and prefer explanations that contrast alternatives [9].

The contributions of the paper are threefold: (1) introducing contrastive highlights, a new local explanation method for highlighting trade-offs between alternative courses of actions that an agent considers, (2) integrating contrastive highlights with reward decomposition to provide users with both a quantification of the expected utility of actions as well as with a visualization of alternative outcomes given a different choice of actions, and (3) conducting users studies showing the integration of both methods results in improved understanding of agent behavior.

2 BACKGROUND

Reward Decomposition: The Hierarchical Reward Architecture (HRA) model, as proposed by Van Seijen et al.[11], receives a decomposed reward function as input and learns a separate Q-function for each reward component. Typically, reward components depend only on a subset of all features therefore, the corresponding Q-function can be approximated by a low-dimensional representation, leading to more effective learning. For a more technical description we refer readers to the original paper [11].

While the reward decomposition approach was originally devised to enable a more efficient learning process, it was suggested by Juozapaitis et al.[6] that it can be used as a local explanation method. Since the individual reward components are mixed into a single reward scalar, in traditional use, Q-values do not give any insight into the positive and negative factors contributing to the agent's decision. However, showing the individual Q-values $Q_c(s, a)$ for each reward component c can explicitly expose the different types of rewards that affect the agent's behavior.

Policy Summaries: Agent Strategy Summarization [2] is an approach for conveying the global behavior of an agent. In this

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.

paradigm, the agent's policy is demonstrated through a carefully selected set of world states. The strategy summarization objective is to select the subset of states that best portrays the policy of the agent. The criteria for selecting states can vary based on the summary objective, e.g., state importance [1, 10] or machine teaching approaches [5, 7].

Building upon this approach, the DISAGREEMENTS algorithm [3] portrays the diverging trajectories of two agents upon reaching a disagreement between them on how best to proceed from a given state. It provides a side-by-side comparison of the difference in outcomes between the agents, constituting a method for agent comparison and behavior difference evaluation. The contrastive highlights method proposed in this paper builds on and extends the DISAGREEMENTS algorithm for the single-agent use case.

3 CONTRASTIVE HIGHLIGHTS

One of the key features of "good" explanations is that they are contrastive [9]. An explanation is contrastive if it provides an answer to the question "Why p rather than q?", where p is the fact which occurred and q is some hypothetical foil which the user might have expected to occur, but did not [8].

We build on the approach of the DISAGREEMENTS algorithm [3], of running and comparing two different agents in parallel, and modify it to instead depict alternative trajectories for a single agent at a given state, each associated with a distinct action available to the agent. These trajectories visualize alternative paths the agent could have taken(*foil q*), had it not chosen the specific action that it had (*fact p*). Relying on prior empirical evidence from the DISAGREEMENTS paper, our method makes use of similar parameters for choosing the summary trajectories.

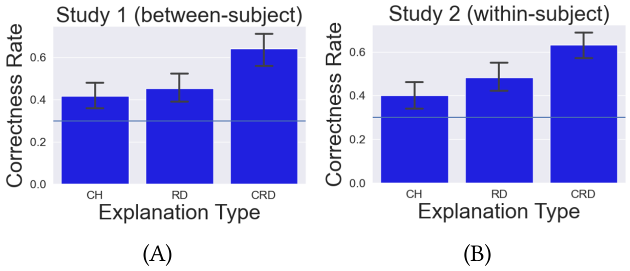
The Algorithm: We initialize and simulate the agent. At each state s_i reached, we note both the agent's preferred and second-best actions. A *contrastive trajectory* is obtained by having the agent initiate the second-best action and progress according to the policy for k steps. The contrastive trajectory is stored and the agent is reverted back to state s_i to progress with its preferred action. upon simulation termination, we pair each state with both the contrastive and true trajectories originating from it. States are then ranked using a chosen importance criteria and the most significant are returned as output.

State Importance: To determine the importance of a state s_i, we compare the two trajectories that branch out of it, *1)* the one chosen (fact p) and *2)* the contrastive (foil q). Importance is then calculated via the *Last-State Importance* metric proposed in [3], which evaluates the significance of the originating state s_i solely based on the last state reached by the compared trajectories. Formally:

$$Im(s_i) = |V(s^p_{i+k}) - V(s^q_{i+k})| \tag{1}$$

Where s^p_{i+k}, s^q_{i+k} denote the states reached by the agent following k steps after selecting the fact(p) and foil(q) in state s_i respectively. This measure utilizes the agent's inherent value function $V(s)$ to describe the estimated utility loss of choosing the foil over the fact (i.e. optimal action). This reflects how "far off" from the original plan the contrastive action has led the agent.

As opposed to DISAGREEMENTS, which only compared conflicting states between the agents, the contrastive highlights algorithm generates a contrastive trajectory at *each* step during execution. While this method does not explicitly answer the original "why" question, it does enable the user to implicitly infer information about the agent's preferences by its choice of action and to observe the short-term alternative outcomes of these.



(A) (B)

Figure 1: Participants' mean success rate in identifying the preferences averaged over all agents by conditions in Study 1 (A) and in Study 2 (B). The error bars show the 95% CI.

Ultimately, the algorithm chooses a limited set of n trajectories to include in the output summary. This distinction between the relevance of states in the overall trace grants it properties of a *global* explanation. However, by changing the algorithms parameters or importance method, it can be tweaked to explicitly provide a local explanation, for instance by depicting the contrastive trajectory for a particular world-state.

4 EXPERIMENTS & RESULTS

To evaluate the benefits of integrating contrastive highlights with reward decomposition as well as their respective contributions to users' understanding of agents' behavior, we conducted two user studies. Participants were presented with explanations in the form of reward decomposition, contrastive highlights, or a combination of both. The explanations were presented for three different agents. Study 1 ($N = 90$) was a between-subjects study, where each participant saw only one type of explanation and for each explanation saw all three agents in random order. Study 2 ($N = 50$) used a within-subjects design. That is, all participants saw all three different agents in a random order, but each agent was accompanied by a different explanation type. Based on these explanations, participants were asked to characterize the reward function of the agent by ranking which of each pair of reward components the agent prioritizes or whether it is indifferent. Additionally, participants answered a 7-point Likert scale explanation-satisfaction questionnaire adapted from [4]. Study 2 participants were also asked to rank their preferences among the three explanation types and describe how each method was helpful in free-text form.

Results: The integration between the two explanation types improved participants' ability to asses the agents' preferences. To measure this ability, we calculated the mean fraction of correct reward component comparisons, i.e., their correctness rate, for each condition. These results were replicated in both studies, and are summarized in Figure 1. For both studies, participants' confidence and satisfaction ratings were above the neutral rating (> 3) but no significant difference between the conditions was observed.

REFERENCES

[1] Dan Amir and Ofra Amir. 2018. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 1168–1176.

[2] Ofra Amir, Finale Doshi-Velez, and David Sarne. 2019. Summarizing agent strategies. *Autonomous Agents and Multi-Agent Systems* 33, 5 (2019), 628–644.

[3] Yotam Amitai and Ofra Amir. 2022. " I Don't Think So": Summarizing Policy Disagreements for Agent Comparison. *Proceedings of the AAAI Conference on Artificial Intelligence* (2022).

[4] 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).

[5] Sandy H Huang, David Held, Pieter Abbeel, and Anca D Dragan. 2019. Enabling robots to communicate their objectives. *Autonomous Robots* 43, 2 (2019), 309–326.

[6] 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*.

[7] Isaac Lage, Daphna Lifschitz, Finale Doshi-Velez, and Ofra Amir. 2019. Exploring computational user models for agent policy summarization. In *IJCAI: proceedings of the conference*, Vol. 28. NIH Public Access, 1401.

[8] Peter Lipton. 1991. The seductive-nomological model: Review of Wesley Salmon Four Decades of Scientific Explanation. *Studies in History and Philosophy of Science Part A* 23, 4 (1991).

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

[10] Pedro Sequeira and Melinda Gervasio. 2020. Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations. *Artificial Intelligence* 288 (2020), 103367.

[11] Harm Van Seijen, Mehdi Fatemi, Joshua Romoff, Romain Laroche, Tavian Barnes, and Jeffrey Tsang. 2017. Hybrid reward architecture for reinforcement learning. *arXiv preprint arXiv:1706.04208* (2017).