Explaining Reinforcement Learning Agents through Counterfactual Action Outcomes

Yotam Amitai*, Yael Septon*, Ofra Amir

Faculty of Data and Decision Sciences, Technion - Israel Institute of Technology
yotama@campus.technion.ac.il, yaelfr1994@gmail.com, oamir@technion.ac.il

Abstract

Explainable reinforcement learning (XRL) methods aim to help elucidate agent policies and decision-making processes. The majority of XRL approaches focus on local explanations, seeking to shed light on the reasons an agent acts the way it does at a specific world state. While such explanations are both useful and necessary, they typically do not portray the outcomes of the agent's selected choice of action. In this work, we propose "COViz", a new local explanation method that visually compares the outcome of an agent's chosen action to a counterfactual one. In contrast to most local explanations that provide state-limited observations of the agent's motivation, our method depicts alternative trajectories the agent could have taken from the given state and their outcomes. We evaluated the usefulness of COViz in supporting people's understanding of agents' preferences and compare it with reward decomposition, a local explanation method that describes an agent's expected utility for different actions by decomposing it into meaningful reward types. Furthermore, we examine the complementary benefits of integrating both methods. Our results show that such integration significantly improved participants' performance.

Introduction

The behavior of reinforcement learning (RL) agents is complex - these agents operate in large, stochastic settings, and are trained through rewards received during their interaction with the environment, which is often sparse. Moreover, state-of-the-art RL models often utilize deep learning, resulting in complex models that can be hard to interpret, making the task of explaining such an agent's behavior even more complicated. While supporting people's understanding of the decision-making of RL agents is challenging, it is important for facilitating effective collaboration and appropriate trust, in particular in high-stakes domains (Rong et al. 2022) such as healthcare or transportation.

Numerous approaches for explainable reinforcement learning (XRL) have been proposed in recent years (Milani et al. 2022). Some methods aim to explain *local* decisions of the agents in particular world-states. These include saliency maps that show what features of the state the agent attends

to (Greydanus et al. 2017), causal action graphs (Madumal et al. 2020b) that are based on a causal model of the agent's policy, or reward decomposition (Juozapaitis et al. 2019) which depict the agent's expected reward for different components of the reward function, among others. In contrast to these local explanation methods, a complementary set of explanation approaches aim to convey the *global* policy of the agent, that is, describe its overall strategy, behavior, or how it behaves in different regions of the state space. Such approaches include policy summaries (Amir, Doshi-Velez, and Sarne 2019) that demonstrate agent behavior in selected world-states based on some criteria or learning a decision tree that approximates the agent's policy (Liu et al. 2018).

In this work, we focus on helping laypeople understand and anticipate agent preferences by depicting the trade-offs between alternative actions. We propose a new method for visualizing the outcomes of actions *not* taken by the agent as a means to explain their chosen action. We introduce "COViz" – **C**ounterfactual **O**utcome **V**isualization[1], a novel local explanation method that looks beyond the agent's decision, to its manifestation. By displaying the outcomes of both actions side-by-side, COViz provides insight into the consequences of the agent's decision. Observing the outcomes ("what if?") can be beneficial for understanding agent behavior or preferences. 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 (Miller 2019).

COViz draws inspiration from global policy summaries (Amir, Doshi-Velez, and Sarne 2019). Similar to policy summaries, COViz conveys agent behavior by showing trajectories of the agent acting in the environment. It is therefore natural to incorporate COViz in policy summaries. We propose a simple extension to COViz for generating global explanations augmented by COViz which selects summary states based on alternative action outcomes.

The idea of presenting action outcomes naturally complements many other local explanation methods that focus on state-limited reasoning. In particular, this approach pairs well with methods that present the agent's *motivation* for choosing a specific action ("why?"), such as *reward decomposition* (Juozapaitis et al. 2019) which quantifies the agent's

---

[1]Code repository: https://github.com/yotamitai/COViz

expectations regarding the utility of different actions.

To evaluate the contribution of COViz to people's understanding of agent preferences, we conducted two user studies in which participants were asked to characterize the reward function of the agent by ranking its preferences. Participants were presented with explanations in the form of counterfactual outcomes (COViz), reward decomposition, or a combination of both approaches. We examined this in the context of multiple different state selection methods, based on global explanations approaches. Both studies showed that combining reward decomposition with COViz improved participants' understanding, compared to showing only one of the explanations. This result suggests that while local explanation methods are typically examined in isolation, their contributions may be complementary, and thus integrating multiple approaches has the potential to benefit users.

Our contributions are fourfold: (1) We introduce COViz, a new local explanation method for highlighting trade-offs between alternative courses of action that an agent considers; (2) We propose a way to generate global explanations augmented by COViz to incorporate alternative outcomes; (3) We integrate COViz with reward decomposition, and (4) We conduct user studies showing both that our new method is beneficial as of itself, and that its integration with reward decomposition improves users' understanding of agent behavior compared to each method separately. Our Supplementary material can be found at https://t.ly/AhrTf

Related Work

Our work draws inspiration from *global* RL explanation approaches. These aim to provide users with a higher-level description of the agent by attempting to portray its behavior, strategy, capabilities, or logic (Huang et al. 2019; Amir, Doshi-Velez, and Sarne 2019; Booth, Muise, and Shah 2019). One approach for conveying the global behavior of an agent is *Agent Strategy Summarization* (Amir, Doshi-Velez, and Sarne 2019). In this 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 state-action pairs that best describes the policy of the agent. The criteria for selecting states can vary based on the summary objective, e.g., state importance (Amir and Amir 2018; Sequeira and Gervasio 2020; Huber et al. 2021) or using machine teaching approaches (Lage et al. 2019; Huang et al. 2019).

Building upon this approach, the DISAGREEMENTS algorithm (Amitai and Amir 2022) 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 COViz method proposed in this paper builds on and extends the DISAGREEMENTS algorithm both for the single-agent use case and for generating a *local* explanation.

Local explanations in XRL seek to describe why a particular choice of action was made by a given agent in a specific world-state (Tabrez, Agrawal, and Hayes 2019; Dodson, Mattei, and Goldsmith 2011), for example by using saliency maps to identify what elements of the environment the agent pays attention to (Hilton et al. 2020; Huber, Schiller, and André 2019; Puri et al. 2020) or generating causal explanations by constructing the agent's action graph (Madumal et al. 2020a). Juozapaitis et al. (2019) use reward decomposition to provide insights into the agents' expected utility by showing users a decomposition of the environment's reward into a sum of meaningful reward types. While this approach shows what the agent expects in terms of utility, it does not show the actual outcomes of its actions. In contrast, COViz demonstrates the tradeoff between actions in terms of trajectories rather than expected reward types.

While growing, there is still an evident lack of research on how to combine different explanation methods. Anderson et al. (2019) presents a user study that investigates the impact of explanations on non-experts' understanding of reinforcement learning agents. They investigate both a common RL visualization, saliency maps, and reward decomposition bars. They designed a four-treatment experiment to compare participants' mental models of an RL agent in a simple Real-Time Strategy game. Their results showed that the combination of both saliency and reward bars was needed to achieve a statistically significant improvement in the mental model score over the control. Other works integrated global explanations in the form of policy summaries with local explanations in the form of saliency maps (Huber et al. 2021) or reward decomposition (Septon et al. 2023). These studies showed that adding reward decomposition explanations improved user performance compared to showing only summaries chosen by HIGHLIGHTS (Amir and Amir 2018) or by frequency-based criteria.

In this work, we introduce a new local explanation method and focus on reward decomposition as a baseline for comparison, since it also illustrates tradeoffs between actions, and because saliency maps were shown to be less helpful (Huber et al. 2021). Other approaches such as causal explanations typically require additional information (e.g., a causal graph) and cannot be implemented directly. Additionally, we show these methods can be combined and why their explanations complement each other.

Background

We assume a Markov Decision Process (MDP) setting. Formally, an MDP is defined by a tuple $< S, A, R, Tr >$ where S is the set of states, A is the set of actions, $R(s, a, s') : S \times A \times S \to \mathbb{R}$ is the reward received after transitioning from state s to state s', due to action a and Tr is a transition probability function $Tr(s, a, s') : S \times A \times S \to [0, 1]$ defining the probability of transitioning to state s' after taking action a in s.

The Q-function is defined as the expected value of taking action a in state s under a policy π throughout an infinite horizon while using a discount factor γ:
$Q^\pi(s, a) = \mathrm{E}^\pi[\sum_{t=0}^{\inf} \gamma^t R_{t+1} | s_t = s, a_t = a]$.

The Highway Environment

In the highway environment (Figure 1) the RL agent controls a green vehicle that navigates a multi-lane highway while

Figure 1: The highway environment. The agent controls the green vehicle.

Local	k	Trajectory length	7
	$Nsim$	Number of simulations	200
	CF_{Meth}	Counterfactual method	SecondBest
Global	n	Summary budget	4
	$overlap$	Max shared states	5
	I_{Meth}	Importance method	Last-State

Table 1: COViz algorithm parameters and study values.

driving alongside and interacting with other (blue) vehicles. The agent can choose to accelerate, decelerate, turn left, turn right or stay idle. The environment can be modified for the number of lanes, vehicle speed range, the number of vehicles, vehicle density, and more. Throughout the paper, we will refer to the lanes from the perspective of the agent, e.g. the bottom-most lane will be denoted as the right-most lane.

A key property of this environment is its lack of an inherent goal that must be reached, instead, an agent's reward is determined by its driving regime. This allows for designing agents with divergent, yet, reasonable strategies as no ground-truth "optimal" behavior exists. Additionally, the environment is both easily understandable to laypeople and represents a relatable real-world scenario.

We focus on a single domain in the current study, as both reward decomposition and approaches similar to COViz (e.g., HIGHLIGHTS, DISAGREEMENTS) have been shown to generalize to additional domains in prior work (Amir and Amir 2018; Juozapaitis et al. 2019; Anderson et al. 2019; Sequeira and Gervasio 2020; Amitai and Amir 2022; Septon et al. 2023).

Reward Decomposition

Hierarchical Reward Architecture (HRA) (Van Seijen et al. 2017) decomposes the agent's reward function into distinct components, each associated with a specific type of reward (e.g., *crashing* or *changing lanes* in the highway environment). This allows the Q-function to be approximated more easily by a low-dimensional representation, which can lead to faster learning. Technically, in the context of deep reinforcement learning, this architecture is implemented in a network that has multiple heads (one for each reward component) that share the lower network layers.

Juozapaitis et al. (2019) used this approach as the basis for a local explanation approach, called *Reward Decomposition*. They present users with Reward Decomposition bars which show the Q-values for each reward component for alternative actions. This aims to reveal the agents' preferences and predictions to users.

Counterfactual Outcomes Visualization

According to cognitive science literature, one of the key features of "good" explanations is that they are contrastive (Miller 2019). Several XRL approaches chose to leverage some form of contrastive information in their explanation techniques (Silva, Bowling, and Lelis 2021; Lin,

Lam, and Fern 2020; van der Waa et al. 2018). Counterfactual explanations are a form of contrastive explanations. A counterfactual explanation 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 (Lipton 1990).

Many explanation techniques are not contrastive and concentrate primarily on the agent's chosen actions (*fact p*), without considering alternative foils. One policy-summarization (global explanation) approach that does provide contrastive information is the DISAGREEMENTS algorithm (Amitai and Amir 2022). It does so by comparing two different agents to one another and visually portraying their points of disagreement.

The DISAGREEMENTS algorithm included an approach for quantifying the importance of a disagreement between the agents at a given state, by comparing their trajectories that branch out of it. The method that proved most efficient for quantifying the differences in these trajectories was comparing the values of the last state (in a limited trajectory) reached by each of them.

We build upon the contrastive nature of DISAGREEMENTS and the idea of focusing on the outcomes of alternative trajectories to devise a local explanation method. Specifically, We adapt the DISAGREEMENTS algorithm to a single agent use case and visually depict trajectories at a given state, each associated with a distinct action available to the agent. In practice, these alternative trajectories are obtained by forcing the agent to take a particular action and then allowing it to progress based on its policy for another k steps. These trajectories visualize different paths the agent *could have* taken (*foil q*), had it *not* chosen the specific action that it had (*fact p*).

We assume that the most relevant counterfactual action for the agent is its second-best choice, as this represents the most likely alternative and often provides the most informative contrast for nonexpert users. While the current focus is on this second-best action, the algorithm can be adapted through the CF_{Meth} parameter (See table 1) to consider other counterfactual options, such as the 'worst' action or a user-selected action, which we acknowledge as a potential area for future investigation.

The COViz algorithm. The pseudo-code for the algorithm is given in Algorithm 1 and parameters are summarized in Table 1 along with their user study values.

The algorithm works as follows: Two empty lists are initialized to account for the execution traces and counterfactual trajectories (lines 3–4). For as many traces as defined, the simulation is initialized and the agent executes its pol-

icy (lines 5–22). At each simulation state s_i, we obtain both the agent's preferred and counterfactual action (lines 9–10). a counterfactual trajectory is obtained by having the agent initiate the counterfactual action and progress according to its policy for another k steps (lines 11–16). The counterfactual trajectory is stored and the agent is reverted back to s_i to progress with its preferred action (lines 17–20). Each state is added to the trace. Once all simulations are completed, traces and counterfactual trajectories are passed to $CFTrajPairs$ where each state is associated with both its true future trajectory and paired to the corresponding counterfactual trajectory, (line 24). The algorithm output is a set of pairs of true and counterfactual trajectories for each state in each trace.

Trajectory length k. The number of steps to include in the counterfactual trajectory (k) was based on the specific environment (highway) implementation, – too few steps result in not enough time to showcase the counterfactual difference while raising the number of steps increases the influence of states succeeding the one being explained.

Complexity. For each state reached by the agent in trace t we generate a counterfactual trajectory of length k. Therefore the complexity is $|t| \times k$, for each trace. Importantly, these computations are done offline (once) and the choice of how many traces to simulate is configurable.

Algorithm 1: Counterfactual Outcome Visualizations

1: **Input:** $\pi, k, Nsims, CF_{Meth}, I_{Meth}, n, overlap$
2: **Output:** CO
3: $Traces \leftarrow$ empty list   *#Execution traces*
4: $C_T \leftarrow$ empty list   *#Counterfactual trajectories*
5: **for** $i = 1$ **to** $Nsims$ **do**
6:    $t \leftarrow$ empty list   *#Current trace*
7:    $sim, s = InitializeSimulation()$
8:    **while** ($!sim.ended()$) **do**
9:       $a^\pi \leftarrow sim.getBestAction(\pi(s))$
10:      $a^C \leftarrow sim.getCFAction(\pi(s), CF_{Meth})$
11:      $c_t \leftarrow$ empty list   *#Counterfactual trajectory*
12:      **for** $j = 1$ **to** k **do**
13:         $s^C \leftarrow sim.advanceState(a^C)$
14:         $a^C \leftarrow sim.getAction(\pi(s^C))$
15:         $c_t.add(s^C)$
16:      **end for**
17:      $C_T.add(c_t)$
18:      $sim, s = reloadSimulation(s)$
19:      $s \leftarrow sim.advanceState(a^\pi)$
20:      $t.add(s)$
21:   **end while**
22:   $Traces.add(t)$
23: **end for**
24: $CO \leftarrow CFTrajPairs(Traces, C_T)$
25:   **Global Extension:**
26: S $\leftarrow topImpTraj(CO, I_{Meth}, n, overlap)$

As opposed to DISAGREEMENTS, which only compared conflicting states between the agents, the COViz algorithm generates a counterfactual trajectory at *each* step during execution. While this method does not explicitly an-swer 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.

Depicting counterfactuals. In terms of visualization, we adopt DISAGREEMENTS's visual comparison approach of displaying the counterfactual agent as a red rectangle originating around the true agent and moving away as their trajectories diverge (see Figure 2). This visualization has been shown to be efficient and satisfying for users (Amitai and Amir 2022).

Global Extension to COViz. The set of counterfactual trajectory pairs obtained by COViz is stored such that a user could ask to view them for any agent-visited state. Alternatively, these can be integrated into a policy summary, showing the most meaningful pairs of counterfactual trajectories. This extension is depicted in Algorithm 1 using the $topImpTraj$ function, (line 26).

After collecting counterfactual trajectories for each of the execution traces' states, their importance is evaluated to generate a ranking. The top-n ranked states will then be used to construct the output summary. 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 counterfactual (foil q). Importance is then calculated via the importance method I_{Meth}. Many criteria can be used to determine trajectory pair importance. We focus on the *Last-State Importance* metric proposed in (Amitai and Amir 2022), which evaluates the significance of the originating state s_i based on the last state reached by the compared trajectories. Formally:

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

Where s^p_{i+k}, s^q_{i+k} denote the states reached by the agent following k steps after selecting the fact(p) or 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 counterfactual action has led the agent.

The global parameter values (Table 1) were chosen based on pilot user studies used to design and optimize the final study. For instance, parameter n which, when raised, increased study duration and cognitive load.

Integrating Counterfactual Outcomes with Reward Decomposition. COViz and reward decomposition naturally complement each other. COViz answers the question "What if?" by depicting alternatives, while reward decomposition answers the question "why?" an action was chosen based on the quantification of the expected utility. For a given state, we can integrate the approaches by showing both the reward decomposition bars for the chosen action and the counterfactual action, as well as presenting the outcome of these choices by showing the trajectories stemming from them. An example of the integration of COViz and reward decomposition can be seen in Figure 2.

Empirical Methodology

To evaluate the benefits of COViz and its integration with reward decomposition, we conducted two user studies asking participants to predict preference differences between agents. Additional experiments were conducted to compare different approaches for state selection for the studies, as further described below.

Agent Training

The agents trained for the experiment were optimized to have preference differences in their driving regimes. Their reward functions specified positive rewards with the following behaviors: changing lanes (CL), high-speed driving (HS), and driving in the right-most lane (RML). All agents received a negative reward for reaching a collision state. To account for each positive reward type, the number of components in the reward decomposition was set to $|C| = 3$.

Three agents were trained with different preferences over the reward types (see Table 2). No future rewards are obtained upon ending the simulation and all agents received the same negative reward for collision. We experimented with different reward-type values to get qualitatively different agent behaviors.

Agents were trained using a double deep Q-Network (DDQN) (Mnih et al. 2015) architecture for 2,000 simulations, each with a maximum of 80 steps. The network input is the observation state represented by an array of size 25 (5X5). The input layer is followed by two fully connected hidden layers of length 256. The final layer connects to three different heads, each of which is designed to account for a specific desired behavior we specified above. Each head consists of a linear layer and outputs a Q-value vector of length of 5 that predicts the Q-value for each of the possible actions: moving left (upwards), idle, moving right (downwards), going faster, going slower. Our implementation is based on two open-source repositories: the Highway environment and its compatible agent architectures (Leurent 2018a,b).

State Selection

To extract the counterfactual outcomes, we ran 200 simulations of each trained agent and saved their traces. States extracted from these traces were used in the study. The main study reported here used the last-state value method described in the COViz global extension section.

In addition, we experimented with two alternative methods for determining state importance (parameter I_{Meth}).

One approach for quantifying a state's importance is to consider the differences in Q-values between different actions (Amir and Amir 2018). This is usually restricted to comparing the agent's preferred action with its second-best or least preferred action. However, this approach proved unsuitable for our studies with COViz, as states selected using this criterion highly incentivized the agents to return to their original path after deviating with a single action, rendering the counterfactual trajectories almost identical to the true ones. We note this effect is domain-dependent – in some

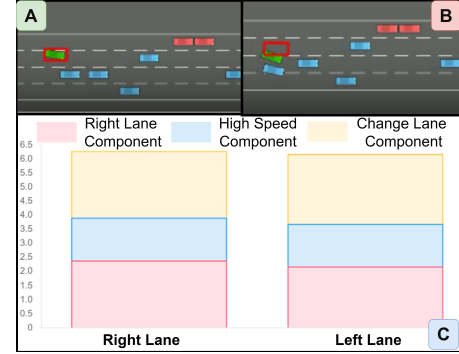| Reward Type	CL	HS	RML	Collision
Agent 1	3	1	8	-3
Agent 2	5	8	1	-3
Agent 3	8	1	5	-3

Table 2: User-study agents' reward-type values.

Figure 2: User study explanation types. (A+B). Counterfactual Outcomes (COViz): videos depicting the different outcomes of counterfactual actions. A and B depict the start and end frame of the video respectively; (C). Reward Decomposition (RD): a graph presenting the value each reward component contributed to the chosen action at the portrayed state. Integrated method (CORD): The COViz outcome video and RD graph together. RD graph portrays both the true and counterfactual action.

domains it is easy to return to the original policy after deviating, while in other domains it may not be possible.

Another approach for selecting states is sampling them based on the frequency of the agent encountering the states, thus reflecting the agent's "typical" behavior. Technically, this can be done by randomly sampling states from execution traces. We ran a replication of our main study with this approach (See supplementary material for complete information, and a short summary in the Results section).

Study Design

We conducted two studies to evaluate the contribution of COViz and its integration with reward decomposition to people's understanding of agents' preferences. Study 1 was a between-subject study, where participants were exposed to a single type of explanation. This design avoids learning effects but suffers from higher variance due to differences between study participants and does not allow to elicit users' subjective preferences. To address this limitation, study 2 used a within-subject design where participants were exposed to all explanation methods, and could thus compare them directly. Both studies were approved by our institutional review board.

Experimental conditions. The studies included three local explanation types: Counterfactual outcomes (*COViz*), reward decomposition (*RD*), and their integration (*CORD*). Examples of explanation types can be seen in Figure 2. Additionally, a control condition of summaries without local

explanations (*SUM*) was tested in study 1. In study 1, participants were assigned a condition and observed all agents using their assigned explanation. In study 2, participants saw each agent with a different, randomly assigned, explanation type. Since it focused on participants' preferences for local explanations, we did not include the summary-only control in this study. The ordering of the *COViz* and *RD* explanations was random, however, the combined method was always shown last as it was more natural to show it after each method appeared separately. While this may introduce a learning effect, study 1 does not suffer from this confound.

Task. In both studies, participants' task was to assess the behavioral preferences of the different agents. For each agent, participants observed four scenarios (i.e., states selected as explained above) and their explanations. Since the Q-values of each agent are different, the "important" states differ as well. However, for a specific agent, all participants viewed the same scenarios. After participants observed the explanation of the agent's behavior in each scenario, they were asked to predict the agent's preferences between two reward components, e.g. "based on the video, which component do you think the agent (green car) cares about more, high-speed or lane change?" (Full surveys in supplementary material). Participants could return and review the explanations while answering the questions. The agents' order and the task answers were randomly presented.

After each task, participants rated their agreement on a 7-point Likert scale (Likert 1932) with statements adapted from Hoffman's explanation satisfaction questionnaire (Hoffman et al. 2018). These adaptations included specifying the task at hand rather than using general terms like 'goal' or 'task', and removing items irrelevant to our study objectives (see supplementary for full scale). Notably, despite these minor modifications, the internal consistency of our adapted scale was confirmed, yielding high Cronbach's Alpha values (>0.9) across all studies.

Then, participants were asked to rate their confidence in each answer on a Likert scale from 1 ("not confident at all") to 5 ("very confident") and they had the option to describe their reasoning in a free-text response. In study 2, where participants saw multiple explanation types, they were also asked to rank their preferences among them and provide textual feedback regarding how each method helped them.

Procedure. In both studies, participants were first introduced to the highway environment. Then, participants learned about their condition's explanation method along with an example to assist in interpreting the visualized information. In study 1, participants were presented with only one of the four explanation types: *Sum*, *COViz*, *RD*, or *CORD*. The explanation type was randomly assigned and consistent across all three tasks. In study 2, before each task, instructions and information regarding the upcoming explanation method were provided, after which, a demonstration for the task they would be asked to complete was shown. Prior to starting the task itself, participants were quizzed on their understanding and only allowed to proceed once answering all questions correctly. Each task included an attention check question. Finally, participants were asked to provide demographic information and rate their proficiency in AI.

All study materials including the specific instructions for each explanation method can be found in the supplementary.

Participants. We recruited participants through Amazon Mechanical Turk ($N_1 = 147$ and $N_2 = 50$ for study 1 and study 2, respectively). On average, study participants self-proclaimed low proficiency in AI which aligns with the purpose of explanations for laypeople. Participants were native English speakers, ages 18 - 55 from the US, UK, or Canada. Participants received a $3 base payment and an additional bonus 30-cent bonus for identifying the preferences of each of the agents correctly. Participants who failed the attention question were excluded from the final analysis, as well as participants who completed the survey in less time than two standard deviations of the mean completion time. After exclusions, we had 115 participants (SUM=27, COViz=30, RD=31, CORD=27) in study 1 and 49 participants in study 2 with a mean age of 39 & 37, and a female count of 39 & 27 for study 1 and 2 respectively. We note that the number of participants in each experiment aligns with previous user studies in XRL, e.g. (Huang et al. 2018; Sreedharan, Srivastava, and Kambhampati 2020).

Results

The objective measure used in our analysis is the mean fraction of correct reward component comparisons, i.e., participants' correctness rate, for each condition. We used the non-parametric Mann-Whitney U test in the statistical analyses. We note that there were no differences between different demographic groups (age or gender).

Overall, the integration of the two explanation types improved participants' ability to asses the agents' preferences. These results were replicated in both studies. In study 1, the integration method of combining both explanation types CORD (*M*= 0.64) led to significantly improved performance compared to RD or COViz alone (RD: M=0.45, RD vs. CORD: $U = 204, p < 0.001$; COViz: M=0.42, COViz vs. CORD: $U = 657.5, p < 0.001$). Comparing to SUM (*M*= 0.31) there was a significant improvement in the performance as well ($U = 607, p = 0.002$). The difference between COViz and RD was not significant ($p = 0.22$). Similarly, in study 2, the combination of explanation types, CORD (*M*= 0.63), significantly improved participants' performance compared to RD or COViz (RD: M=0.48, RD vs. CORD: $U = 767, p < 0.001$ and COViz: M= 0.4, COViz vs. CORD: $U = 1814.5, p < 0.001$). Here too, the difference between COViz and RD was not significant ($p = 0.064$).

Moreover, in our experimentation with frequency-based state selection, we found that all local explanation methods were useful, with no significant difference between them, when addressing informative states. However, for less informative states all explanation methods proved less useful than a coin toss, except the integrated approach which was still superior (see supplementary material).

While CORD participants were subjected to more information than the other conditions, prior work has demonstrated that more information does not necessarily lead to better outcomes (Anderson et al. 2019), implying that integration between these methods successfully captures rele-
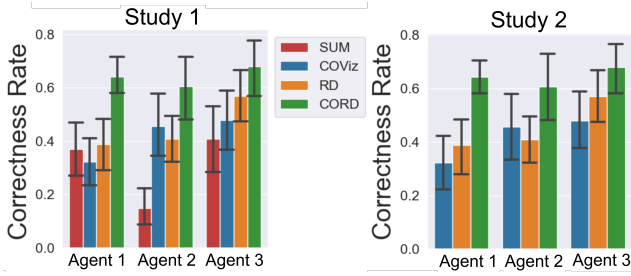
Figure 3: Participants' mean success rate in identifying the preferences by condition and agent in study 1 (Left) and study 2 (Right). The error bars show the 95% CI.

Figure 4: The distribution of participants' ranking of the three explanation types

vant and complementary information for performance without imposing excessive cognitive load.

In study 1, when breaking down the results by agent, the trend was consistent with the overall results. The greatest improvement from the integration of the two approaches was for Agent 1 (Figure 3). We note that Agent 1 is a tricky agent as its second preference - being in the right-most lane is correlated with its first preference changing lanes. This shows that the combination of the two explanations has a significant impact, especially for challenging behaviors. Participants' comments also mentioned the benefit of combining the different explanations. For instance, one participant wrote "The two [explanations] together helped me understand but separately I was a lot more confused". Another comment was "The graphs help to visualize how the agent is rewarded for behaviors while the videos show the results of rewarding (or prioritizing) those actions for the agent."

In study 2, we asked participants to rank the three explanation types. Participants differed in their preferences. Some preferred the video-based approach of COViz, e.g., "video is always easy for me to comprehend", while others preferred the reward bars: "The graph is more efficient". A participant that preferred the combination wrote: "It is more instructive to both see the component values and then watch the video. " The overall ranking results show that COViz was ranked first by 28 participants (57% of participants), while the combination of the two explanation types was ranked in last place by 26 participants (53% of participants) as shown in Figure 4.

For both studies, participants' confidence and satisfaction ratings were above the neutral rating (> 3). Specifically, for study 1 and study 2 the satisfaction results for each of the conditions were COViz: M=4.9 and 5.16, RD: M=5.06 and 5.06, CORD: M=5.29 and 5.04 respectively and SUM=5.4 in study 1 (Graphs provided in supplementary). With respect to participants' confidence, the ratings by condition observed in studies 1 and 2 were COViz: M=4.04 and 3.93, RD: M=3.8 and 4.02, CORD: M=4.03 and 3.83, respectively). None of the differences between conditions in terms of satisfaction or confidence were significant.

Discussion and Future Work

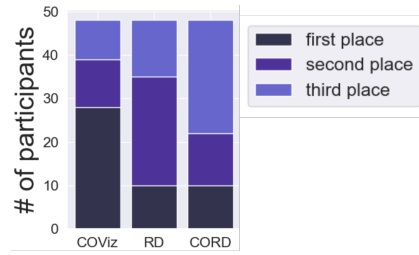This paper presented a new method for explaining RL agents through counterfactual outcomes, and the benefits of combining them with another local explanation - reward decomposition. We conducted user studies to evaluate the contribution of this approach to people's ability to analyze agent preferences as well as rating the participants' preferences of each explanation type. Our results show that both explanation types, COViz, and reward decomposition, resulted in better-than-random participant performance, but the integration of the two enabled participants to reach a significantly higher success rate. It is not trivial that this would be the case since the benefit might only stem from one explanation, and the cognitive load might be higher as was observed in some studies (Anderson et al. 2020). Indeed, when showing frequency-based chosen states, COViz or reward decomposition alone did not always result in higher performance than random guess. However, showing both types of explanations together did improve performance beyond random guessing.

Even though the combination of the explanation types led to a higher correctness rate, when participants were asked to rate the explanation types, the combination was ranked last. We hypothesize that the reason for that is the mental overload on the participants. For example, one participant commented that "...the combination is too confusing for me." When combining two explanation types participants are required to pay attention to more details. This finding is in line with previous work stating that subjective satisfaction and proxy tasks often do not align with objective performance (Buçinca et al. 2020). A potential future-work solution is to reduce user cognitive load by cascading the information and allowing users to control the amount they view at once, thus avoiding overwhelming them and engaging them more.

Our experimentation of state-selection criteria highlighted a vulnerability of COViz – only a single counterfactual action is constrained. In the generalized continuous setting, a single action deviation does not necessarily result in a divergent outcome and is more likely to be negligible or easily reversible. Future work can alleviate this limitation by enforcing an alternative sequence of actions rather than forcing only a single deviation from the policy.

Additionally, the counterfactual actions were chosen heuristically in this work, however, it is also possible to delegate this decision to the user to increase interactivity and trust-building or for debugging purposes.

Finally, while this work focused on the task of identifying agent preferences, future work can try to characterize which explanation types are most useful for different user tasks, and explore the combination of other local explanation types, including saliency maps, action graphs, and others.

Acknowledgements

References

Amir, D.; and Amir, O. 2018. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 1168–1176.

Amir, O.; Doshi-Velez, F.; and Sarne, D. 2019. Summarizing agent strategies. *Autonomous Agents and Multi-Agent Systems*, 33(5): 628–644.

Amitai, Y.; and Amir, O. 2022. " I Don't Think So": Summarizing Policy Disagreements for Agent Comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Anderson, A.; Dodge, J.; Sadarangani, A.; Juozapaitis, Z.; Newman, E.; Irvine, J.; Chattopadhyay, S.; Fern, A.; and Burnett, M. 2019. Explaining reinforcement learning to mere mortals: An empirical study. *arXiv preprint arXiv:1903.09708*.

Anderson, A.; Dodge, J.; Sadarangani, A.; Juozapaitis, Z.; Newman, E.; Irvine, J.; Chattopadhyay, S.; Olson, M.; Fern, A.; and Burnett, M. 2020. Mental models of mere mortals with explanations of reinforcement learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(2): 1–37.

Booth, S.; Muise, C.; and Shah, J. 2019. Evaluating the Interpretability of the Knowledge Compilation Map: Communicating Logical Statements Effectively. In *IJCAI*, 5801–5807.

Buçinca, Z.; Lin, P.; Gajos, K. Z.; and Glassman, E. L. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, 454–464.

Dodson, T.; Mattei, N.; and Goldsmith, J. 2011. A natural language argumentation interface for explanation generation in Markov decision processes. In *International Conference on Algorithmic Decision Theory*, 42–55. Springer.

Greydanus, S.; Koul, A.; Dodge, J.; and Fern, A. 2017. Visualizing and Understanding Atari Agents. *arXiv preprint arXiv:1711.00138*.

Hilton, J.; Cammarata, N.; Carter, S.; Goh, G.; and Olah, C. 2020. Understanding RL Vision. *Distill*, 5(11): e29.

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

Huang, S. H.; Bhatia, K.; Abbeel, P.; and Dragan, A. D. 2018. Establishing appropriate trust via critical states. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3929–3936. IEEE.

Huang, S. H.; Held, D.; Abbeel, P.; and Dragan, A. D. 2019. Enabling robots to communicate their objectives. *Autonomous Robots*, 43(2): 309–326.

Huber, T.; Schiller, D.; and André, E. 2019. Enhancing Explainability of Deep Reinforcement Learning Through Selective Layer-Wise Relevance Propagation. In Benzmüller, C.; and Stuckenschmidt, H., eds., *KI 2019: Advances in Artificial Intelligence*, 188–202. Cham: Springer International Publishing. ISBN 978-3-030-30179-8.

Huber, T.; Weitz, K.; André, E.; and Amir, O. 2021. Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence*, 301: 103571.

Juozapaitis, Z.; Koul, A.; Fern, A.; Erwig, M.; and Doshi-Velez, F. 2019. Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on Explainable Artificial Intelligence*.

Lage, I.; Lifschitz, D.; Doshi-Velez, F.; and Amir, O. 2019. Exploring computational user models for agent policy summarization. In *IJCAI: proceedings of the conference*, volume 28, 1401. NIH Public Access.

Leurent, E. 2018a. An Environment for Autonomous Driving Decision-Making.

Leurent, E. 2018b. rl-agents: Implementations of Reinforcement Learning algorithms.

Likert, R. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Lin, Z.; Lam, K.-H.; and Fern, A. 2020. Contrastive explanations for reinforcement learning via embedded self predictions. *arXiv preprint arXiv:2010.05180*.

Lipton, P. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27: 247–266.

Liu, G.; Schulte, O.; Zhu, W.; and Li, Q. 2018. Toward interpretable deep reinforcement learning with linear model u-trees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 414–429. Springer.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2020a. Distal Explanations for Model-free Explainable Reinforcement Learning. *arXiv preprint arXiv:2001.10284*.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2020b. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2493–2500.

Milani, S.; Topin, N.; Veloso, M.; and Fang, F. 2022. A Survey of Explainable Reinforcement Learning. *arXiv preprint arXiv:2202.08434*.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533.

Puri, N.; Verma, S.; Gupta, P.; Kayastha, D.; Deshmukh, S.; Krishnamurthy, B.; and Singh, S. 2020. Explain Your Move: Understanding Agent Actions Using Specific and Relevant Feature Attribution. In *8th International Conference on Learning Representations, ICLR*. OpenReview.net.

Rong, Y.; Leemann, T.; trang Nguyen, T.; Fiedler, L.; Seidel, T.; Kasneci, G.; and Kasneci, E. 2022. Towards Human-centered Explainable AI: User Studies for Model Explanations. .

Septon, Y.; Huber, T.; André, E.; and Amir, O. 2023. Integrating policy summaries with reward decomposition for explaining reinforcement learning agents. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, 320–332. Springer.

Sequeira, P.; and Gervasio, M. 2020. Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations. *Artificial Intelligence*, 288: 103367.

Silva, C. R.; Bowling, M.; and Lelis, L. H. 2021. Teaching People by Justifying Tree Search Decisions: An Empirical Study in Curling. *Journal of Artificial Intelligence Research*, 72: 1083–1102.

Sreedharan, S.; Srivastava, S.; and Kambhampati, S. 2020. Tldr: Policy summarization for factored ssp problems using temporal abstractions. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 30, 272–280.

Tabrez, A.; Agrawal, S.; and Hayes, B. 2019. Explanation-based reward coaching to improve human performance via reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 249–257. IEEE.

van der Waa, J.; van Diggelen, J.; Bosch, K. v. d.; and Neerincx, M. 2018. Contrastive explanations for reinforcement learning in terms of expected consequences. *arXiv preprint arXiv:1807.08706*.

Van Seijen, H.; Fatemi, M.; Romoff, J.; Laroche, R.; Barnes, T.; and Tsang, J. 2017. Hybrid reward architecture for reinforcement learning. *arXiv preprint arXiv:1706.04208*.