OPEN ACCESS

University of BRISTOL

Liu, X., McAreavey, K., & Liu, W. (2023). Contrastive Visual Explanations for Reinforcement Learning via Counterfactual Rewards. In L. Longo (Ed.), *Explainable Artificial Intelligence: First World Conference, xAI 2023, Lisbon, Portugal, July 26–28, 2023, Proceedings, Part II* (1 ed., Vol. 2, pp. 72-87). (Communications in Computer and Information Science; Vol. 1902). Springer. https://doi.org/10.1007/978-3-031-44067-0_4

Peer reviewed version

License (if available):
CC BY

Link to published version (if available):
10.1007/978-3-031-44067-0_4

Link to publication record in Explore Bristol Research
PDF-document

This is the accepted author manuscript (AAM). The final published version (version of record) is available online via Springer at https://doi.org/10.1007/978-3-031-44067-0_4 . Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research
General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

Contrastive Visual Explanations for Reinforcement Learning via Counterfactual Rewards

Xiaowei Liu, Kevin McAreavey, and Weiru Liu

School of Engineering Mathematics and Technology,
University of Bristol, UK
{xiaowei.liu,kevin.mcareavey,weiru.liu}@bristol.ac.uk

Abstract. Causal attribution aided by counterfactual reasoning is recognised as a key feature of human explanation. In this paper we propose a post-hoc contrastive explanation framework for reinforcement learning (RL) based on comparing learned policies under actual environmental rewards vs. hypothetical (counterfactual) rewards. The framework provides policy-level explanations by accessing learned Q-functions and identifying intersecting critical states. Global explanations are generated to summarise policy behaviour through the visualisation of sub-trajectories based on these states, while local explanations are based on the action-values in states. We conduct experiments on several grid-world examples. Our results show that it is possible to explain the difference between learned policies based on Q-functions. This demonstrates the potential for more informed human decision-making when deploying policies and highlights the possibility of developing further XAI techniques in RL.

Keywords: Explainable reinforcement learning · Contrastive explanations · Counterfactuals · Visual explanations.

1 Introduction

The aim of explainable AI planning (XAIP) and explainable reinforcement learning (XRL) is to help end-users better understand agent behaviour (e.g. learned policies) and how that behaviour relates to the environment (i.e. transition probabilities and rewards) [15,12,6]. Contrastive explanations are a particular approach to explainable AI (XAI) that seek to answer *contrastive why-questions*, with the aim of identifying the causes of one event (called the fact) relative to the causes of another (called the foil in the counterfactual case, meaning that the event did not occur in the actual world) [26]. Miller [31,32] emphasised the importance of contrastive explanations in explainable AI (XAI) based a survey of the relevant literature from philosophy and social science. Many recent studies have explored different aspects of contrastive explanations in XAIP and XRL [49,16,35].

One possibility for contrastive explanations in XRL is to compare a learned policy under actual environmental rewards versus a learned policy under hypothetical (counterfactual) rewards. Such comparisons have analogies in several

areas of RL. For example, preference-based RL [8,20,27] seeks to learn a policy that is optimal with respect to altered rewards that combine environmental rewards with human preferences. If a policy is learned under both kinds of rewards, then it opens the possibility of explaining one policy with respect to the other by way of contrast. An interesting research challenge then is how to generate contrastive explanations for RL to help humans better understand the impact of actual rewards on learned agent behaviour.

In this paper, we develop a framework for contrastive explanations in RL that compares the policy learned under actual rewards against policies learned under different counterfactual rewards. The actual reward configuration is just the actual rewards, while each counterfactual reward configuration is a partial alteration of the actual rewards. We assume that all policies are otherwise trained under the same conditions (e.g. same hyperparameters, same training steps). We adopt a post-hoc XAI paradigm to provide two types of contrastive explanation:

1. *Global explanation*: This type of explanation focuses on providing overall policy explanations about an agent's behaviour. It provides insights into how these policies behave in general by visualising (sub-)trajectories, and how decisions are made in some states among the configurations.
2. *Local explanation*: This type of explanation addresses the question, "Why was action a chosen in state s rather than action a'?" It provides more fine-grained information based on the action-value function in each configuration, allowing for a better understanding of agent behaviour.

The rest of this paper is organised as follows. Section 2 reviews related literature about explanation in XAIP and XRL. Section 3 formulates the main structure of constrastive explanation, and Section 4 offers illustrative explanation and further analysis on the cases. The last section offers conclusions, discussions and future works.

2 Related Work

Explainable AI (XAI) has obtained significant attention in recent years, driven by the advancement and wide application of machine learning and AI systems especially in decision making [23,44,40]. The systems pose challenges for trustworthiness if they simply employ more powerful and flexible models, albeit at the expense of model interpretability and transparency [33,12,30]. The complexity of the systems, as well as the difficulty explaining an agent's behaviour in planning and RL, have been acknowledged by many research papers [55,5,6] which further assessed the necessity of XAI for planning and RL. In this part, we review some literature that is closely related to the topics in XRL.

Policy summarisation in RL. Policy summarisation has been a subject of much research in XAIP and XRL [24,45], which improves interpretability and provides an explanation regarding the agent's policy behaviour. One approach is the use of trajectory visualisation, which involves summarising the agent's policy by extracting important trajectories from simulations. For example, in

[1], the authors discussed the design and implementation of the *HIGHLIGHTS* algorithm, which used state importance and the state diversity criteria for choosing the trajectories from the replay buffer. This approach was further extended in [18], which integrated saliency maps to local explanation through the visualisation of trajectories. In robotics and control, [17] utilises example trajectories to enable users to better anticipate the behaviours or goals of robots. Following this, [24] enhanced the example trajectories extraction by optimising an inverse reinforcement learning or imitation learning problem. Another approach to policy summarisation is generating an abstracted or hierarchical explanation through learned models or data about the policy. For instance, in [47], authors generated policy-level explanations for RL, which used a Markov chain to represent abstracted states and their transitions based on the training data. In [43], authors proposed a framework for learning hierarchical policies in multi-task RL that can learn human instructions and generate an explanation of its decisions by learned instructions back to humans. Similarly, in [54], authors proposed a policy abstraction method through an extended model of MDP for deep Q-networks. Besides, many prior studies have demonstrated effectiveness revealing an agent behaviour through trajectory visualisation and policy abstraction [3,34,19]. These works provide solid support for trajectory visualisation that serves as an effective approach to policy summarsation and explaining the agent's behaviour. Building upon this foundation, we extend these methods by incorporating contrastive explanations.

Critical states and key moments for explanation in RL. [16] suggested that the essence of the policy relies on a few critical states or the corresponding agent's actions on those states, and proposed approaches for computing critical states based on the action-value function and the policy function. Similarly, [22] explored the importance of a state with the variance of its learning action-value function on states. Another study by [41] proposed a method which extracted key moments of the agent's decision with statistical information of the agents, delivered visual summaries and offered user studies of the performance. The authors further extracted key elements of interestingness from an agent's learning experience in [42], and presented a global and visual summarisation of agent behaviour based on elements including frequency and sequence. From another aspect, counterfactual state, which was proposed in [36] captured the key states that an agent chose a different action with minimal change to the input of the policy networks. Deep generative models were used to create counterfactual states and present visual counterfactual explanations to users on Atari games in this work. Recent research integrated generating counterfactuals in latent space with gradient-driven methods [53]. In the domain of robust RL, the detection of critical states against adversarial attacks adopted this metric [25]. Other studies [54,11,13] focused on the identification and visualisation of the salience of state features for Atari agents, which could be considered a metric of critical states.

Explanation via rewards or value functions in RL. Notably, the contrasting descriptions were provided for users' queries related to predefined state transitions and expected reward outcomes of the agent [49]. This approach did not di-

rectly answer the contrastive questions on the agent's behaviour, but transformed the questions and provided answers by explaining the learned value functions instead. Similar to [14], the proposed method introduced contrastive explanations regarding the simulated outcomes of the rollouts based on two policies (the agent policy and the foil policy). The construction of the fact and foil in these papers, and the scheme for contrastive explanation are heuristics, which partially motivated the contrastive explanation for the difference in reward configurations in our work. The framework in [10] provided a policy evaluation method on the action-value function that identified the influence of state transitions by removing some transition data. According to [29], contrastive explanations were generated by action influence models which involved causal relationship of rewards and actions. [21] introduced an explanation framework based on reward decomposition, in which it is assumed that rewards can be decomposed into vector-like rewards with semantic meaning. It is extended in a user-study for real-time strategy games in [2], generated explanations for outcomes that agents intended to achieve in tabular RL approaches [52]. [28] further utilised reward decomposition to build a learnable framework for robotics.

From a boarder aspect of XRL, some works have considered aspects of user needs, such as personalised explanations [46] and the complexity of contrastiveness [35]. We refer readers to see systematic overview of topics in XRL [37,51,56,48].

3 Generating Contrastive Explanations for Two Policies

3.1 Preliminaries

In this work we consider infinite-horizon, discounted reward Markov Decision Processes (MDPs) [38,39]. An MDP is a tuple $\mathcal{M} = (S, A, P, R, \gamma)$ where S is a finite set of states, A is a finite set of actions, $P : S \times A \to \Delta(S)$ is a (stochastic) transition function where $\Delta(S)$ is the set of probability distributions over S, $R : S \times A \times S \to \mathbb{R}$ is a reward function, and $\gamma \in [0, 1)$ is a discount factor. The transition function P says if action a is executed in state s then the system will transition to state s' with probability $P(s, a, s')$, where $P(s, a, s')$ denotes the probability of reaching state s' according to distribution $P(s, a)$. The optimal value function $V^\star$ is defined for each $s \in S$ as:

$$V^\star(s) = \max_{a \in A} \sum_{s' \in S} P(s, a, s') \left[R(s, a, s') + \gamma V^\star(s')\right] \tag{1}$$

and the optimal action-value function $Q^\star$ is defined for each $a \in A$ as:

$$Q^\star(s, a) = \sum_{s' \in S} P(s, a, s') \left[R(s, a, s') + \gamma V^\star(s')\right] \tag{2}$$

A policy is a function $\pi : S \to A$. The optimal policy $\pi^\star$ can be extracted directly from the optimal action-value function, i.e. for each $s \in S$:

$$\pi^\star(s) = \text{argmax}_{a \in A} Q^\star(s, a) \tag{3}$$

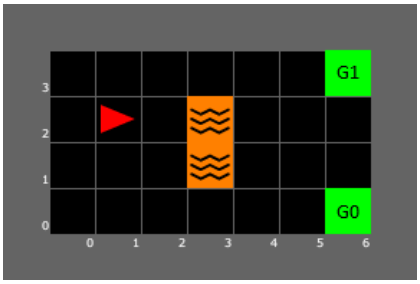


Fig. 1: A grid-world scenario.

In planning (where P and R are known) a well-known approach to finding the optimal value function is value iteration [4]. In RL (where P and R are unknown) a well-known approach to finding the optimal action-value function is Q-learning [50]. In our proposed method, we assume access to both the learned policy and the learned action-value function as optimal functions defined in Equation (2) and (3), which allows us to generate explanations from the decision-making processes of the agent. We consider MDPs with different reward functions R_i as $\mathcal{M}_i = (S, A, P, R_i, \gamma)$. The optimal policy, optimal value function and optimal action-value function on $\mathcal{M}_i$ are marked as $\pi_i^\star, V_i^\star(s)$ and $Q_i^\star(s, a)$.

Environment description: A demo of Grid-World. We consider a simple case with a 7×4 grid-world (Figure 1). Four actions, *UP, DOWN, LEFT, RIGHT*, are available at each state with a random action rate with 0.1.[1] To reach the final destinations (*G0* and *G1* in green blocks) with the same positive reward, the agent (red triangle) has to avoid the absorbing states, the lava cells (orange), with a reward of 0. The agent initialises at one of the four cells on the far-left side of the lava, and every action taken receives a penalty of -0.01.

3.2 Identifying Critical States from Q-functions

Critical states are defined as states where small changes can significantly affect the agent's behaviour, and they have been shown to be reliable indicators of an agent's decision-making process [16]. One of the most commonly used metrics for defining critical states is the difference between the maximum and average action values of a state above a predetermined threshold. Let $\mathcal{C}_i$ denote the set of critical states under the optimal policy $\pi^\star$ for a given MDP. We refer this metric as *Max-mean* [16],

$$\mathcal{C}_i = \left\{ s \in S \mid \left(\max Q_i^\star(s, a) - \frac{1}{|A|} \sum_a Q_i^\star(s, a) \right) > \tau \right\}. \tag{4}$$

[1] With 90% probability the agent moves one cell in the direction specified by the action (i.e. the action succeeds), or with 5% probability each the agent moves one cell either clockwise or anti-clockwise relative to the direction specified by the action (i.e. the action fails). This grid-world was implemented by Minigrid[7].

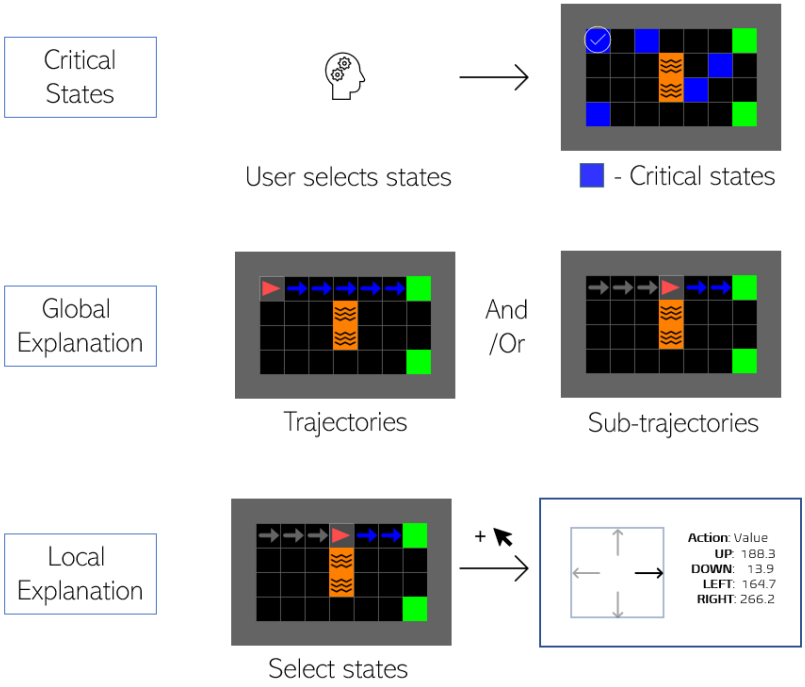
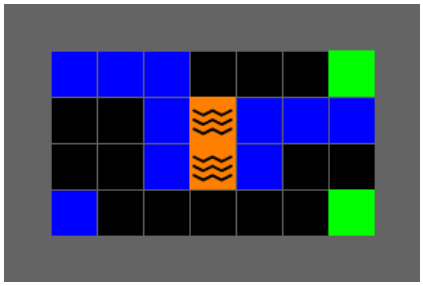


Fig. 2: An illustration of explanation process. Global explanation: agent (red triangle) starts at a state, an example (full) trajectory and/or sub-trajectory are visualised. Local explanation: local explanations are provided with a window on states of interest by interacting with the explainer, and with more information explained on the agent's state.

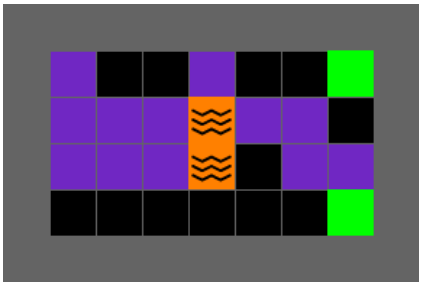
The number of critical states can vary depending on the reward function of the MDP. By changing the threshold τ according to the user's needs and the environmental reward function, the number of critical states can be adjusted accordingly. If there are K MDPs, we can denote the set of intersected states among these MDPs as $\mathcal{C}^{\mathrm{I}} = \cap_{i=1}^{K} \mathcal{C}_i$.

One of the commonly used metrics for the max-mean metric is the difference between the maximum and minimum action-values from the action-value function [1]. Another study by [22] explores the importance of a state by examining the variance of its action-value function at states during learning. We consider these as variants of the Max-mean approach. We acknowledge that further evaluation of these methods through user studies is necessary to determine their efficacy in generating useful explanations of agents. A survey of related work on critical states and key moments is provided in Section 2, and we offered analysis in Section 5.3.

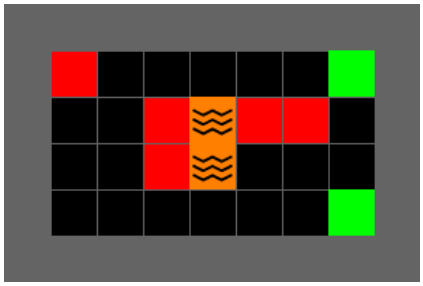
Before presenting more details, we provide an overview of how our methodology (referred to as the *explainer*) generates explanations for users (referred to as



(a) Critical states (blue) with one reward function.



(b) Critical states (purple) with a different reward function.



(c) The intersected critical states (red) of two reward functions.

Fig. 3: Critical states from two different reward functions, and the intersected critical states which hint the important states in common of the two configurations.

the *explainee*). The explainer initiates the process by generating critical states based on the specific questions of the explainee. Critical states are generated as a series of intersecting critical states if there are multiple policies. These critical states are then represented visually as contrastive trajectories. Each trajectory records the sequence of state-action pairs an agent takes, beginning from each critical state during simulations. Additionally, to provide further details to the user, the explainee can pause the visualisation and inquire about states of interest. In the event of such queries, the explainer presents contrastive explanations on different learned policies, including the feasible actions that can be executed, the relevant action values from those states, along with the optimal actions of each policy.

3.3 Global Explanation by Using Critical States

Firstly, the explainer presents a number of states based on a default threshold τ of the metric in Equation (4). These states are then used as inputs to the explainer, and Monte Carlo simulations are initiated in parallel, recording the state-action pairs until termination states are reached (i.e., absorbing states or

predefined maximum length of recording). We refer a *full trajectory* as a trajectory rollout history in which agent starts from the initial state of the environment and terminates until the agent reaches termination states. A *sub-trajectory* is a trajectory rollout history in which agent starts from the critical states and reaches termination states. To provide a comprehensive global explanation on states, we visualise full trajectories or partial trajectories rollouts (illustrated in Figure 2). Finally, the corresponding trajectories with the maximal probability for the counterfactual reward function within the sample space are presented to the user either as videos or images with all the state-action pairs highlighted in contrast. These trajectories serve as contrastive global explanations, allowing the explainee to observe, comprehend the agent's behaviour and compare agents with respect to their reward functions in each configuration.

3.4 Local Explanation and Contrastive Explanation Based on Action-values

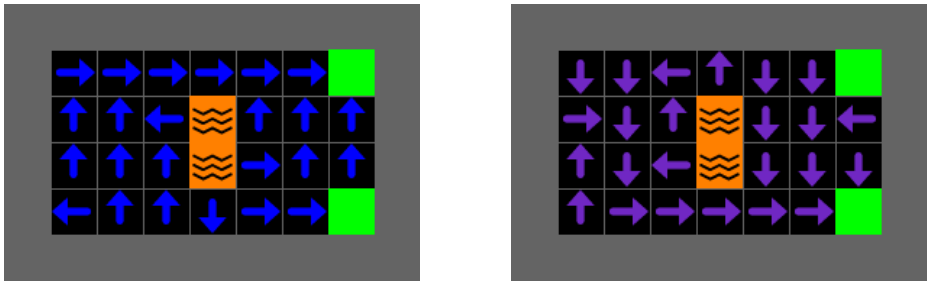
If the users have further queries regarding how the policy acts on specific states, we visualise based on the states in question by displaying optimal actions and action-values of those states. We leverage the learned action-value function to generate local explanations for the agent's decision-making. For instance as shown in Figure 5a and 5d, the action *RIGHT* is the optimal action as the explainee observe that it has the highest value. The explainer displays the relative importance of each action at a given state based on its action-value, and provides a more interpretable and informative explanation for the agent's decision.

We provide contrastive explanations on critical states in each reward configuration, highlighting the differences between the learned policies and their corresponding action-values. Specifically, we contrastively display the different critical states presented in the reward configuration based on the metric in Equation (4). The intersected critical states are highlighted (for instance, in red in Figure 3) to draw the attention of explainees to the potential significance of the states across multiple configurations. In our proposed framework, the explainee can choose specific states of interest, and the explainer will then display all the actions taken by agents and the action-values pairs from agents in a contrastive manner across different reward configurations. This allows the explainee to observe the different action-values pairs associated with the same action, and possible different optimal actions in a given state for better understanding of agents' behaviour. In addition, we can further enhance the local explanations by considering the uncertainty of the agent's action-value estimation.

4 Experiments

We consider two variants of this grid-world named GW^{+} and GW^{-} (see in Figure 1) where the the reward functions are set as:

– GW^+: The agent will receive a reward of +1 at *G0* (6,4) and a reward of +3 at *G1* (6,0).
– GW^-: The agent will receive a reward of +3 at *G0* (6,4) and a reward of +1 at *G1* (6,0).



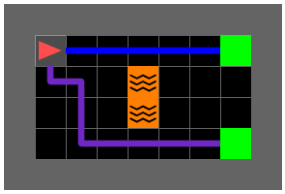
(a) A learned policy in GW^+. (b) A learned policy in GW^-.

Fig. 4: Illustration of grid-world by Q-learning in GW^+ and GW^- (blue and purple).

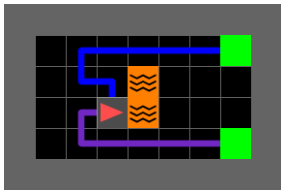
The purpose of the setting is to give an illustrative example where reward functions are the only part vary and the transition functions in the MDPs remain the same. We designed such two intuitive reward functions under which Q-learning is used to learn policies. Specifically, we ran the algorithm on two grid-worlds denoted as GW^+ and GW^-, respectively, with a discount factor of 0.99 and learning rate of 0.01. The Q-tables are initialised with values $\mathcal{N}(0, 1)$ and 14000 episodes. After the training process we output the Q-table as the learned action-value function.

We identify the critical states from each Q table and compute their intersection set $\mathcal{C}^{\mathrm{I}}$, which provides a simple illustration of policy behaviour. To compute the critical states, we utilised the Max-mean method in Equation (4) and set a predetermined parameter of $\tau = 80$ for better illustration. The resulting critical states for GW^+ and GW^- are shown in Figure 3. There were five intersected critical states, and we selected three of them for illustration: (0,3), (2,1), and (4,2).

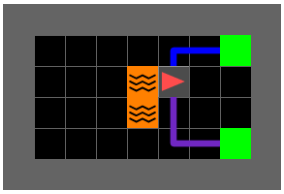
To provide a global explanation, we report the learned policies for GW^+ and GW^- in Figure 4a and Figure 4b, respectively, along with the optimal actions at each state indicated by arrows. We then present further global and contrastive explanations based on a sample of simulations shown in blue and purple colours in Figure 5a, 5b, and 5c. The corresponding states are highlighted in the images. The explainees can observe that the agent's decisions starting from certain states can lead to completely different goal states which reveals the importance of understanding critical states and their impact on the overall policy behaviour. For example, in Figure 5b, this figure illustrates two trajectories which are legible to the explainees in the presence of two possible goal states of the agent and the



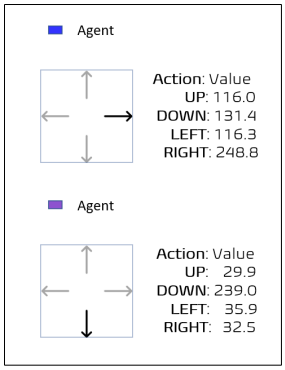
(a) Sub-trajectory from (0,3).



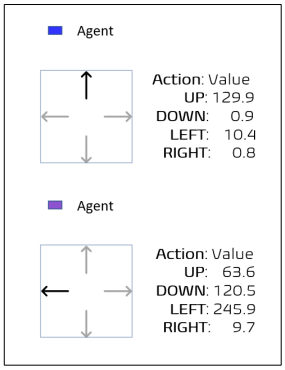
(b) Sub-trajectory from (2,1).



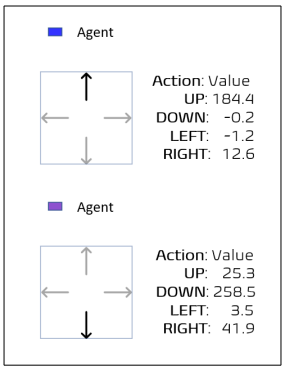
(c) Sub-trajectory from (4,2).



(d) Local information on (0,3).



(e) Local information on (2,1).



(f) Local information on (4,2).

Fig. 5: An example of constrasive explanation on three critical states. 5a-5c: Global contrastive explanation on agents in GW^+ (blue) and GW^- (purple). 5d-5f: Local contrastive explanation on actions in GW^+ and GW^-.

avoidance of lava states. If the agent starts at the left position next to the lava grid, with one policy, it takes the action *UP* and *LEFT*, then executed a series of action of *RIGHT* and eventually reaches the goal state on the top-right (*G1*). With a different policy, it takes the action *LEFT* and *DOWN*, and then provides another series of actions that reaches the goal state of bottom-right (*G0*). We observe at least two agent behaviours: the behaviour of reaching different goal states, and the behaviour of stepping away from the lava grid. From the perspective of the explainees' mental models, we wish they would attribute causes of the difference of reward configurations themselves from these behaviours, with possibly further observation on the local explanation of action-values. Though the visual explanation does not directly tell the explainees the actual factors on how and why the reward differs, it illustrates an explicable trajectory that help them comprehend the objectives of the agent's behaviours. Similar explanations apply to Figure 5a and Figure 5c.

Local and contrastive explanation are shown in Figure 5d, 5e, and 5f. Providing action-values and optimal actions for each state in contrast contributes to a more comprehensive explanation of global contrastive explanation with trajectories. The explainee can observe differences among the action-values across

actions, which could help explainees comprehend why the agent chose the learned action (highlighted in black) over the other three actions (shown in grey). For instance, in Figure 5e, it has been demonstrated that the action *UP* is the optimal action for one agent, the blue agent, as it yields the highest action-value of 129.9. On the other hand, for the second agent, the purple agent, the optimal action is *LEFT*, with a corresponding optimal action-value of 245.9. The difference in optimal actions aid the explainees in attributing causal factors, e.g., why the agent ultimately reaches distinct goal states.

5 Conclusion and Discussion

In this study, we addressed the problem of explanation in RL by comparing policies based on their action-value functions where the policies are learned under different reward functions. Our proposed methods generating global and local explanations through trajectories based on intersected critical states. We further showed our explanation successfully demonstrating the contrastive behaviour by an example from Q-learning in a grid-world.

5.1 Discussion of Research Questions

In this subsection, we discuss our research questions and the knowledge contributed to the XAI community in this paper. The utilisation of counterfactual rewards within XRL is to address two broader and significant research questions:

Research Question 1: Casual attribution via counterfactual reasoning. Suppose that an action X has been learned by an agent and the explainees asked "why X?" as the action may look unexpected or weird. Humans are believed to answer such questions by identifying causes through counterfactual reasoning. In RL, the learned action in each state depends on characteristics of the underlying MDP, which consists of a transition function and a reward function. A reasonable cause in RL then might reference characteristics of the transition function and/or reward function that led to action X having been learned. An important question would be: what characteristics of the transition function and/or reward function are most relevant to the action X having been learned? In this paper, we limit our focus to the reward function. The objective of simulating hypothetical rewards is not to imply that X would not have been learned in the absence of the actual reward function. Instead, its purpose is to facilitate counterfactual reasoning in humans, enabling them to attribute characteristics of the actual reward function as causes for X having been learned. We focus on predefined hypothetical reward functions, but our objective remains the same: as an aid to understand the actual reward function and its impact on learned actions.

Research Question 2: Casual contrastive explanations. Suppose Research Question 1 has been answered and the explainees are able to attribute characteristics of the actual reward function as causes for X having been learned. Suppose again that the explainees proposed some other action Y which would

have been normal/expected, and asked "why X rather than Y?" According to the question, action Y was not learned by the agent, so the explainees are not able to attribute characteristics of the actual reward function as causes for Y having been learned. Instead, we need a hypothetical reward function, and specifically one where Y would have been learned with all the settings being equal (i.e., the same transition function, hyper-parameters, and training steps etc.) However, if we have those causes, then we can answer the question by focusing on the aspects where the actual and hypothetical causes differ. In the paper we do not directly address Research Question 2, but we do lay some groundwork on how it could be addressed, mainly due to the need to construct hypothetical reward functions, but also in the need for visual comparisons. However, a major difference is that the hypothetical reward function is now significant; it must ensure that Y is learned, all else being equal. The same criteria may be reasonable for choosing hypothetical reward functions under Research Question 1.

5.2 Discussion of Findings

This paper contributions to the field of XRL in the sense that it addresses a previously unexplored question improving the users' comprehension of the agent behaviour through the construction of a hypothetical reward function. Specifically, we use the learned policies on both the hypothetical and actual reward functions to enable users to engage in counterfactual reasoning on the discrepancies existed between these reward functions. The proposed method offers a viable and natural means of addressing contrastive questions and limit the information scope to identification of critical states and trajectory visualisation. The metric used for critical states in this study builds upon a prior research. The visualisations presented in this paper leverage the established groundwork of trajectory visualisation methods, which have proven to be an effective approach to summarising policies and an agent's behaviour. We emphasise the importance of co-use for explaining the difference of reward functions: contrastive explanations visually based on trajectories and utilisation of action-values.

5.3 Limitations and Future Work

While this paper primarily focuses on computational methods rather than user studies, it is important to acknowledge the need for a user study to evaluate the effectiveness of the visual explanations provided and the validity under specific conditions. We recognise the significance of conducting a comprehensive user study as part of future work. We also provide possible future improvement on the following topics.

Critical states identification. One limitation observed is the absence of user evaluation regarding the metrics employed for critical states identification. While the action-values can reveal the optimal action(s) that are preferred over alternative actions, future work should focus on providing explanations from the underlying reasons supporting such preferences, e.g., an epistemic perspective of certainty/uncertainty of the agent. Furthermore, in addition to computing

critical states based on the action-value or value function, we posit that a similar metric can be applied to the policy function and potentially extended to continuous action spaces.

Textual explanation and interactive interface. The proposed method primarily provide visual comparisons to facilitate casual attribution by humans, however, this could fail when the visualisation does not meet human's expectation. We recognise this limitation, and textual-based and question-based explanations could be used in enhancing the potential cognitive process by explainees in future work. The inclusion of an interactive interface is targeted to consider the needs and preferences in explanation for users [35,46]. For instance, providing users with the capability to specify the desired number of critical states or certain type of metric they wish to view, particularly in situations where there may be an overwhelming number of states to consider. Moreover, particular attention would be given to prioritising the presentation of trajectory explanations that involve disagreement perceived by the explainees.

System design. The proposed method exhibits limitations when applied to complex environments. The method heavily relies on an accurate model or simulator to generate trajectories supposing that the agent can be positioned in arbitrary states. Alternative solutions would be to compute critical states through pre-recording trajectories or employing episodic memory of an agent [9] in future work. While the computational cost increases when multiple policies need to be trained for real-world applications, the training of contrastive policies can be conducted in parallel. And in most scenarios, we believe that a limited form of contrastive explanations can be achieved sufficiently with only two policies. Furthermore, exploring the explanation of potential policy randomness and environmental uncertainty (e.g., random effects and transitions induced by the environment or random actions taken by the agent) is identified as a promising future direction.

Acknowledgement

The authors would thank anonymous reviewers for their valuable comments. This work is partially funded by the EPSRC CHAI project (EP/T026820/1).

References

1. Amir, D., Amir, O.: Highlights: Summarizing agent behavior to people. In: AAMAS'18. p. 1168–1176 (2018)
2. Anderson, A., Dodge, J., Sadarangani, A., Juozapaitis, Z., Newman, E., Irvine, J., Chattopadhyay, S., Fern, A., Burnett, M.M.: Explaining reinforcement learning to mere mortals: An empirical study. In: IJCAI'19. p. 1328–1334 (2019)
3. Annasamy, R., Sycara, K.: Towards better interpretability in deep q-networks. In: AAAI'19. vol. 33, pp. 4561–4569 (2019)
4. Bellman, R.E.: Dynamic programming. Princeton university press (2010)
5. Chakraborti, T., Kulkarni, A., Sreedharan, S., Smith, D.E., Kambhampati, S.: Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. In: ICAPS'19. vol. 29, pp. 86–96 (2019)
6. Chakraborti, T., Sreedharan, S., Kambhampati, S.: The emerging landscape of explainable automated planning & decision making. In: IJCAI'20. pp. 4803–4811 (2020), survey track
7. Chevalier-Boisvert, M., Willems, L., Pal, S.: Minimalistic gridworld environment for gymnasium (2018), https://github.com/Farama-Foundation/Minigrid
8. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: NeurIPS'17. vol. 30 (2017)
9. Cruz, F., Dazeley, R., Vamplew, P.: Memory-based explainable reinforcement learning. In: AI 2019: Advances in Artificial Intelligence. pp. 66–77. Cham (2019)
10. Gottesman, O., Futoma, J., Liu, Y., Parbhoo, S., Celi, L., Brunskill, E., Doshi-Velez, F.: Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. In: ICML'20. vol. 119, pp. 3658–3667 (2020)
11. Greydanus, S., Koul, A., Dodge, J., Fern, A.: Visualizing and understanding atari agents. In: ICML'18. pp. 2877–2886 (2018)
12. Gunning, D.: Darpa's explainable artificial intelligence (xai) program. Proceedings of the 24th International Conference on Intelligent User Interfaces p. ii (2019)
13. Gupta, P., Puri, N., Verma, S., Kayastha, D., Deshmukh, S., Krishnamurthy, B., Singh, S.: Explain your move: Understanding agent actions using specific and relevant feature attribution. In: ICLR'20 (2020)
14. Hayes, B., Shah, J.A.: Improving robot controller transparency through autonomous policy explanation. In: 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 303–312 (2017)
15. Hoffmann, J., Magazzeni, D.: Explainable AI Planning (XAIP): Overview and the Case of Contrastive Explanation (Extended Abstract), pp. 277–282. Springer International Publishing, Cham (2019)
16. Huang, S.H., Bhatia, K., Abbeel, P., Dragan, A.D.: Establishing appropriate trust via critical states. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3929–3936. IEEE (2018)
17. Huang, S.H., Held, D., Abbeel, P., Dragan, A.D.: Enabling robots to communicate their objectives. Autonomous Robots **43**, 309–326 (2017)
18. Huber, T., Weitz, K., André, E., Amir, O.: Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. Artificial Intelligence **301**, 103571 (2021)
19. Hüyük, A., Jarrett, D., Tekin, C., van der Schaar, M.: Explaining by imitating: Understanding decisions by interpretable policy learning. In: ICLR'21 (2021)

20. Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., Amodei, D.: Reward learning from human preferences and demonstrations in atari. NeurIPS'18 **31** (2018)
21. Juozapaitis, Z., Koul, A., Fern, A., Erwig, M., Doshi-Velez, F.: Explainable reinforcement learning via reward decomposition. In: Arxiv (2019)
22. Karino, I., Ohmura, Y., Kuniyoshi, Y.: Identifying critical states by the action-based variance of expected return. In: ICANN'20. pp. 366–378. Springer (2020)
23. Kober, J., Bagnell, J.A., Peters, J.: Reinforcement learning in robotics: A survey. The International Journal of Robotics Research **32**, 1238 – 1274 (2013)
24. Lage, I., Lifschitz, D., Doshi-Velez, F., Amir, O.: Exploring computational user models for agent policy summarization. In: IJCAI'19. p. 1401–1407 (2019)
25. Lin, Y.C., Hong, Z.W., Liao, Y.H., Shih, M.L., Liu, M.Y., Sun, M.: Tactics of adversarial attack on deep reinforcement learning agents. In: IJCAI'17. p. 3756–3762 (2017)
26. Lipton, P., Knowles, D.: Contrastive Explanations, p. 247–266. Royal Institute of Philosophy Supplements, Cambridge University Press (1991)
27. Liu, R., Bai, F., Du, Y., Yang, Y.: Meta-reward-net: Implicitly differentiable reward learning for preference-based reinforcement learning. NeurIPS'22 **35**, 22270–22284 (2022)
28. Lu, W., Magg, S., Zhao, X., Gromniak, M., Wermter, S.: A closer look at reward decomposition for high-level robotic explanations. ArXiv **abs/2304.12958** (2023)
29. Madumal, P., Miller, T., Sonenberg, L., Vetere, F.: Explainable reinforcement learning through a causal lens. In: AAAI'20. pp. 2493–2500 (2020)
30. Marcus, G., Davis, E.: Rebooting AI: Building Artificial Intelligence We Can Trust. Pantheon Books, USA (2019)
31. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
32. Miller, T.: Contrastive explanation: a structural-model approach. The Knowledge Engineering Review **36**, e14 (2021)
33. Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. ArXiv **abs/1706.07979** (2017)
34. Mott, A., Zoran, D., Chrzanowski, M., Wierstra, D., Rezende, D.J.: Towards interpretable reinforcement learning using attention augmented agents. In: NeurIPS'19. p. 12360–12369 (2019)
35. Narayanan, S., Lage, I., Doshi-Velez, F.: (when) are contrastive explanations of reinforcement learning helpful? ArXiv **abs/2211.07719** (2022)
36. Olson, M.L., Khanna, R., Neal, L., Li, F., Wong, W.K.: Counterfactual state explanations for reinforcement learning agents via generative deep learning. Artificial Intelligence **295**, 103455 (2021)
37. Puiutta, E., Veith, E.M.S.P.: Explainable reinforcement learning: A survey. ArXiv **abs/2005.06247** (2020)
38. Puterman, M.L.: Markov decision processes: Discrete stochastic dynamic programming. In: Wiley Series in Probability and Statistics (1994)
39. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach (4th Edition). Pearson (2020)
40. Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T.P., Silver, D.: Mastering atari, go, chess and shogi by planning with a learned model. Nature **588**, 604 – 609 (2019)
41. Sequeira, P., Gervasio, M.: Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations. Artificial Intelligence **288**, 103367 (2020)

42. Sequeira, P., Hostetler, J., Gervasio, M.T.: Global and local analysis of interestingness for competency-aware deep reinforcement learning. ArXiv **abs/2211.06376** (2022)
43. Shu, T., Xiong, C., Socher, R.: Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. In: ICLR'18 (2018)
44. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T.P., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D.: Mastering the game of go without human knowledge. Nature **550**, 354–359 (2017)
45. Sreedharan, S., Srivastava, S., Kambhampati, S.: Tldr: Policy summarization for factored ssp problems using temporal abstractions. ICAPS'20 **30**, 272–280 (2020)
46. Sreedharan, S., Srivastava, S., Kambhampati, S.: Using state abstractions to compute personalized contrastive explanations for ai agent behavior. Artificial Intelligence **301**, 103570 (2021)
47. Topin, N., Veloso, M.: Generation of policy-level explanations for reinforcement learning. In: AAAI'19. pp. 2514–2521 (2019)
48. Vouros, G.A.: Explainable deep reinforcement learning: State of the art and challenges. ACM Comput. Surv. **55**(5) (dec 2022)
49. Waa, J., Diggelen, J., Bosch, K., Neerincx, M.: Contrastive explanations for reinforcement learning in terms of expected consequences. In: IJCAI 2018 - Explainable Artificial Intelligence (XAI) Workshop (2018)
50. Watkins, C.J., Dayan, P.: Q-learning. Machine learning **8**, 279–292 (1992)
51. Wells, L., Bednarz, T.: Explainable ai and reinforcement learning—a systematic review of current approaches and trends. Frontiers in Artificial Intelligence **4** (2021)
52. Yau, H., Russell, C., Hadfield, S.: What did you think would happen? explaining agent behaviour through intended outcomes. NeurIPS'20 **33**, 18375–18386 (2020)
53. Yeh, E., Sequeira, P., Hostetler, J., Gervasio, M.T.: Outcome-guided counterfactuals for reinforcement learning agents from a jointly trained generative latent space. ArXiv **abs/2207.07710** (2022)
54. Zahavy, T., Ben-Zrihem, N., Mannor, S.: Graying the black box: Understanding dqns. In: ICML'16. pp. 1899–1908 (2016)
55. Zelvelder, A.E., Westberg, M., Främling, K.: Assessing explainability in reinforcement learning. In: Explainable and Transparent AI and Multi-Agent Systems. pp. 223–240. Springer International Publishing (2021)
56. Čyras, K., Rago, A., Albini, E., Baroni, P., Toni, F.: Argumentative xai: A survey. In: IJCAI'21. pp. 4392–4399 (2021), survey Track