Counterfactual Explanations for Reinforcement Learning Agents

Doctoral Consortium

Jasmina Gajcin
Trinity College Dublin
Dublin, Ireland
gajcinj@tcd.ie

ABSTRACT

Reinforcement learning (RL) algorithms often use neural networks to represent agent's policy, making them difficult to interpret. Counterfactual explanations are human-friendly explanations which offer users actionable advice on how to change their features to obtain a desired output from a black-box model. However, methods for generating counterfactuals in RL ignore the stochastic and sequential nature of RL tasks, and can generate counterfactuals which are difficult to obtain, affecting user effort and trust. My dissertation focuses on developing methods that take into account the complexities of RL framework and provide counterfactual explanations that are easy to reach and confidently produce the desired output.

KEYWORDS

Reinforcement Learning; Explainability; Counterfactual Explanations; Contrastive Explanations; Causality

ACM Reference Format:
Jasmina Gajcin. 2023. Counterfactual Explanations for Reinforcement Learning Agents: Doctoral Consortium. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

In the recent years, reinforcement learning (RL) algorithms have achieved remarkable success in numerous tasks, and are being developed for high-risk applications, such as autonomous vehicles [16]. However, RL algorithms often rely on neural networks to represent the agent's policy, making their behavior difficult to understand and interpret [18]. As RL agents are developed for real-life tasks, understanding their behavior is necessary for ensuring user trust and encouraging human-AI collaboration.

In the earlier stages of my PhD research I explored contrastive (Section 2) and causal explanations (Section 3), before deciding on counterfactual explanations as the research topic (Section 4).

2 CONTRASTIVE EXPLANATIONS FOR COMPARING AGENTS' PREFERENCES

The majority of XRL approaches focus on explaining one policy [1, 10, 13, 18]. However, understanding the differences between policies is necessary to understand how different reward functions affect agent's behavior or in situations where user needs to choose between multiple policies. To that end, we proposed an algorithm for comparing and explaining the preferences of RL agents trained on different reward functions [8]. Our approach generates contrastive explanations about two policies π_A and π_B by analysing the state space in which policies disagree on the best course of action due to the difference in their preference. Our approach can differentiate between differences in behavior that stem from different abilities of policies and those caused by the different preferences of equally capable agents. We then use only data on preference-based differences of policies to generate contrastive explanations that describe state features that individual policies favor. We evaluated our approach in a merging task for autonomous vehicles, where we compare a speed-oriented with a safety-oriented policy.

3 RECCOVER: DETECTING CAUSAL CONFUSION FOR EXPLAINABLE RL

Causal confusion occurs when agent relies on the spurious correlations between features that might not hold across the state space [5]. If an agent is deployed to an environment where such correlation is broken, its decisions are misguided and performance suffers.

We proposed ReCCoVER (Recognizing Causal Confusion for Verifiable and Explainable RL) [7], an algorithm for detecting and correcting causal confusion in critical states in RL agents. While previous work can detect causal confusion only after performance drops in states where spurious correlations are broken [5, 12], ReCCoVER detects and corrects causal confusion before deployment. ReCCoVER detects causal confusion in a specific state by testing agent's performance in alternative environments where spurious correlations might not hold. Alternative environments are generated by performing causal interventions on state features, which can break correlation between them. Additionally, a feature-parametrized policy π_G is trained to simultaneously learn a separate policy for each feature subset, and evaluated in the alternative environments to uncover whether ignoring certain features during training leads to more robust policies. Causal confusion is detected when agent's performance decreases significantly in an alternative environment, but a policy relying on a subset of features G' performs well. ReCCoVER then advises the developers to only rely on the features from G' in the critical state s.

Although aimed at developers of RL systems, ReCCoVER offers actionable advice on how to change the feature space to avoid causal confusion, making it a human-friendly explanation method. Additionally, ReCCoVER uncovers causal relationships in agent's reasoning, which is inherent to human way of understanding events.

4 COUNTERFACTUAL EXPLANATIONS IN RL

The main research topic of my thesis is developing counterfactual explanations for RL agents. Counterfactuals interpret the decisions

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

of black-box models by answering the question: *"Given that black-box model outputs A for input features $f_1, \cdots f_k$, how can the features be changed so that output B is obtained?"* [22]. For example, if a person's loan application is rejected by an automated system a counterfactual explanation could help them understand how they can change their features in order to be approved in the future. Counterfactual explanation of an instance x is given in the form of a counterfactual instance x', which is as similar as possible to x but produces the desired output. Counterfactuals are a natural extension to my earlier PhD work, as they are actionable, contrastive and causal.

As they can suggest life-altering changes to user's features, counterfactual explanations carry a great responsibility. Counterfactuals that are difficult to obtain or do not deliver the desired outcome can cost users time and effort, which can cause frustration and decrease trust in the system. For that reason, counterfactual explanations are often evaluated against counterfactual properties,in order to to find those which are most suitable for the user. For example, validity is evaluated as the probability that the counterfactual instance produces the desired outcome, and proximity is a feature-based similarity measure used to choose the counterfactual that is most similar to the original instance. Similarly, sparsity measures the number of feature changes between the original instance and the counterfactual, and data manifold closeness ensures that the counterfactual instance falls within the realm of realistic instances, to ensure it's easily obtainable [21].

Numerous methods have been developed for generating counterfactual explanations for supervised learning tasks [3, 14, 17, 22]. Search for counterfactuals often consists of defining a loss function comprising multiple counterfactual properties and optimizing it over the training data set. The approaches differ in their definition of the loss function and the choice of the optimization approach. In RL counterfactuals have been used to explain the choice of action A in state with features f_1, ..., f_k by generating a similar counterfactual state where model chooses a different action B. The only approach to generating counterfactuals in RL relies on a similar approach as supervised learning methods and uses generative models to find counterfactuals that are similar in features to the original state [15].

In my first work on exploring counterfactual explanations in RL, we conducted a survey of the state-of-the-art counterfactual approaches in supervised and RL and analysed the main differences between the two frameworks from the perspective of counterfactual explanations [6]. One of the main findings of this research is that the idea of easily obtainable instance varies significantly between supervised and RL. While supervised learning deals with one-step prediction tasks, RL focuses on sequential and often stochastic tasks. This means that in RL two states can have similar features, but be far from each other in terms of execution. Relying solely on the feature-based counterfactual properties, while sufficient in supervised learning, can generate counterfactuals that are difficult to obtain or do not deliver the desired outcome in RL tasks. Additionally, while supervised learning models make predictions based only on input features, decisions of RL agents are motivated by a wider range of causes, such as goals, objectives or outside events. To fully understand agent's behavior, counterfactuals should not rely only on state features, but include all potential causes of a decision [4].

As a result of this research, I devised a research question which is guiding my current and future work:

> How can counterfactual explanations be redefined to account for the complex, sequential and stochastic nature of RL tasks?

While our work in Gajcin and Dusparic [6] offers theoretical justification for redefining counterfactuals for RL tasks, we are currently working on implementing the first algorithm for generating RL-specific counterfactual explanations. We have redefined and implemented novel counterfactual properties from RL perspective and use them to guide the search for counterfactuals. Specifically, we look for counterfactuals that optimize the following properties:

(1) *Reachability:* it is possible for two states to be similar in features but far away in terms of execution. We define reachability as the minimum number of RL actions necessary to navigate from the original to the counterfactual state.
(2) *Cost-efficiency:* while current work assumes there is no difference in cost of changing different features, in reality, some changes require more effort. We define cost-efficiency as the minimal cumulative RL cost of performing actions to transform the original into the counterfactual state.
(3) *Stochastic certainty:* during the process of transforming the original into the counterfactual state, stochasticity in the environment can affect the state. We define stochastic certainty as the probability of obtaining the desired outcome after performing the sequence of actions that transforms the original into the counterfactual state.

Currently, we are implementing an approach for generating counterfactuals that optimize the above properties. We aim to explore if relying on RL-specific counterfactual properties can generate more easily obtainable counterfactuals compared to methods optimizing traditional, feature-based counterfactual properties. Moreover, we are conducting a user study to explore how RL-specific counterfactual explanations affect user understanding of agents.

Evaluation is one of the biggest obstacles to developing counterfactual explanations. While supervised learning approaches are often evaluated on the same datasets (e.g. German credit [3, 14], Breast Cancer [2, 11] or MNIST dataset [9, 11, 19, 20]), in RL there is not established benchmark for evaluating counterfactuals. Additionally, few works evaluate counterfactuals in a user study [15], despite counterfactuals being heralded as human-friendly explanations. A user-focused investigation of the best ways to present and evaluate counterfactuals will make a part of my remaining PhD work.

In the future work, we plan to explore the personalization aspect of counterfactual generation. Namely, counterfactual search optimizes different, often conflicting objectives, which might not all be equally important to the user. We hope to explore how human-in-the-loop approaches can be applied to counterfactual search to ensure that the explanations align with user's preferences. Furthermore, we plan to investigate ways for generating counterfactuals that do not rely only on changing state features, but also goals, objectives and other causes of decisions specific to the RL framework.

ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of a grant from Science Foundation Ireland under Grant number 18/CRT/6223. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

REFERENCES

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

[2] Ziheng Chen, Fabrizio Silvestri, Gabriele Tolomei, He Zhu, Jia Wang, and Hongshik Ahn. 2021. ReLACE: Reinforcement Learning Agent for Counterfactual Explanations of Arbitrary Predictive Models. *arXiv preprint arXiv:2110.11960* (2021).

[3] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. 2020. Multi-objective counterfactual explanations. In *International Conference on Parallel Problem Solving from Nature*. Springer, 448–469.

[4] Richard Dazeley, Peter Vamplew, and Francisco Cruz. 2021. Explainable reinforcement learning for Broad-XAI: a conceptual framework and survey. *arXiv preprint arXiv:2108.09003* (2021).

[5] Pim De Haan, Dinesh Jayaraman, and Sergey Levine. 2019. Causal confusion in imitation learning. *Advances in Neural Information Processing Systems* 32 (2019).

[6] Jasmina Gajcin and Ivana Dusparic. 2022. Counterfactual Explanations for Reinforcement Learning. *arXiv preprint arXiv:2210.11846* (2022).

[7] Jasmina Gajcin and Ivana Dusparic. 2022. ReCCoVER: Detecting Causal Confusion for Explainable Reinforcement Learning. In *Explainable and Transparent AI and Multi-Agent Systems*, Davide Calvaresi, Amro Najjar, Michael Winikoff, and Kary Främling (Eds.). Springer International Publishing, Cham, 38–56.

[8] Jasmina Gajcin, Rahul Nair, Tejaswini Pedapati, Radu Marinescu, Elizabeth Daly, and Ivana Dusparic. 2021. Contrastive Explanations for Comparing Preferences of Reinforcement Learning Agents. *arXiv preprint arXiv:2112.09462* (2021).

[9] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2017. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443* (2017).

[10] Guiliang Liu, Oliver Schulte, Wang Zhu, and Qingcan Li. 2018. Toward interpretable deep reinforcement learning with linear model u-trees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 414–429.

[11] Arnaud Van Looveren and Janis Klaise. 2021. Interpretable counterfactual explanations guided by prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 650–665.

[12] Clare Lyle, Amy Zhang, Minqi Jiang, Joelle Pineau, and Yarin Gal. 2021. Resolving Causal Confusion in Reinforcement Learning via Robust Exploration. In *Self-Supervision for Reinforcement Learning Workshop-ICLR 2021*.

[13] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 2493–2500.

[14] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 607–617.

[15] Matthew L Olson, Roli Khanna, Lawrence Neal, Fuxin Li, and Weng-Keen Wong. 2021. Counterfactual state explanations for reinforcement learning agents via generative deep learning. *Artificial Intelligence* 295 (2021), 103455.

[16] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S Iyengar. 2018. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 1–36.

[17] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. FACE: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 344–350.

[18] Erika Puiutta and Eric Veith. 2020. Explainable reinforcement learning: A survey. In *International cross-domain conference for machine learning and knowledge extraction*. Springer, 77–95.

[19] Robert-Florian Samoilescu, Arnaud Van Looveren, and Janis Klaise. 2021. Model-agnostic and Scalable Counterfactual Explanations via Reinforcement Learning. *arXiv preprint arXiv:2106.02597* (2021).

[20] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2019. Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *arXiv preprint arXiv:1905.07857* (2019).

[21] Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596* (2020).

[22] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.