Explainable Multi-Objective Reinforcement Learning

Juan C. Rosero[*]

School of Computer and Statistics, Trinity College Dublin, Ireland

1 Introduction

Reinforcement Learning (RL) is a machine learning paradigm that focuses on trial-and-error, where the agent interacts with the environment and receives rewards and penalties for its actions [21]. It has achieved significant breakthroughs across multiple domains such as video games, vehicle guidance, system control, robotics, networking, and healthcare [7]. However, real-world problems often require optimizing for multiple, potentially conflicting objectives, rather than a single scalar reward, which has led to the development of Multi-Objective Reinforcement Learning (MORL), where agents learn to balance and reason over several competing goals simultaneously. Recent work has demonstrated the potential of MORL for tackling complex decision-making scenarios, including more traditional autonomous systems [6] or even more specific applications like exploring and optimizing pandemic mitigation policies [19]. But, while MORL is a more realistic and flexible modeling paradigm, it also increases the complexity of both learning and decision-making.

On the other hand Explainable Reinforcement Learning (XRL) has also emerged to address growing demands for transparency, trust, and human-in-the-loop interaction in RL systems. XRL surveys [17, 14] generally classify methods into inherently explainable approaches and post-hoc explanations, aiming to provide insight into the decision-making of RL agents. However, most existing XRL work focuses on single-objective RL and overlooks the added challenges of explaining decisions in multi-objective settings.

The intersection of these two fields remains largely underexplored. Multi-objective scenarios intuitively introduce additional complexity for decision making, as agents must balance trade-offs, adapt to evolving priorities, and reason about interactions between competing goals. Existing XRL methods do not provide insight into how such trade-offs are managed or how decisions are influenced by evolving preferences or by the balancing of objectives. That is the gap I aim to explore and address in my research.

2 Explainable Multi-objective Reinforcement Learning

To understand the research gaps presented by the intersection of MORL and XRL, Explainable Multi-Objective Reinforcement Learning (XMORL), we first briefly introduce both fields.

[*] Corresponding Author. Email: roserolj@tcd.ie

2.1 Multi-Objective Reinforcement Learning

The common approaches to MORL can be classified in single-policy methods, which produce one policy methods that attempts to balance all objectives, and multi-policy methods, that try to generate a set of policies that reflect different trade-offs or preferences. Single-policy approaches, such as those using weighted sums or utility functions [1, 15], are simple but often obscure the trade-offs involved. In contrast, multi-policy approaches like Deep W-Learning (DWN) [8] and Pareto-Conditioned Networks (PCN) [19] provide more flexibility while modeling the problems and behaviors, but add complexity to understanding agent behavior and decision-making because of the addition of a new layer to the decision-making process to select one policy at each moment.

2.2 Explainable Reinforcement Learning

XRL methods are usually classified by either scope, in either global or local explanations, or the timing of the explanation, in intrinsic or post-hoc explanations [17]. Local explanations focus on individual decisions, while global methods describe overall policy behavior. Intrinsic methods build interpretability into the agent itself, while post-hoc approaches are applied after training. Examples of local intrinsic methods include reward decomposition [11] and HexRL [16]. Global intrinsic methods, such as SkillTree [23] and CSG [22], aim to structure the agent's policy in such a way that makes the overall behavior easier to understand. Local post-hoc methods, like counterfactual explanations [5] or DeepSHAP [24], explain specific actions after they occur. Finally, global post-hoc methods, such as policy summarization [2] and decision tree interpretations [12], produce interpretable overviews of learned policies without modifying the agent.

3 Research Questions

The focus of my research is how to make decision-making in MORL more transparent and interpretable. MORL enables decision making in scenarios where single-objective rewards are not enough to effectively model the desired behavior of an agent, but it also complicates the decision process by introducing an additional layer of complexity that makes agent behavior harder to explain to users. To address this problem, I focus on the field of Explainable Multi-Objective Reinforcement Learning (XMORL). Based on this, I proposed the following research questions:

1. How can explanations be designed to accommodate objective-specific reasoning and trade-offs in multiobjective settings? An interesting approach is to combine per-objective reasoning with a global visualization of explanation of the trade-offs involved.

2. How can XRL tools be improved to accommodate dynamic and evolving preferences? Enhancing preference-conditioned MORL methods like PCN [18] to be more easily explainable seems to be a possible approach to address this.
3. Can explanation methods be made scalable to high-dimensional objective spaces, where trade-offs become increasingly complex? One possible approach to this problem could be to explore hierarchical or modular approaches.
4. How can user-in-the-loop XRL frameworks be adapted to support preference and trade-off refinement in MORL? To address this problem it is required to develop frameworks that not only help explain behavior of the agent, but also provide mechanisms to understand and shape objective prioritization.
5. Can causal explanations be implemented in MORL so explanations can explain the perceived trade-offs and the underlying reasoning behind them? The starting step would be to extend existing causal XRL approaches [13] to work in a multi-objective setting.
6. Which benchmarks and evaluation mechanisms are required to evaluate performance and quality on XMORL methods? Developing benchmarks and standardized quality metrics would be critical to effectively compare and improve different approaches. While existing XRL metrics provide a starting point, they may not directly translate to the multi-objective aspects of XMORL. For example, commonly used metrics like coverage may require adaptation, as it's not clear if they would need to be assessed separately for the state and objective space or not.

These questions aim to bridge the current gap between MORL and XRL, supporting the development of trustworthy, explainable, and adaptive MORL applications.

4 PhD Work Plan

Our research focuses on the intersection of MORL and XRL, we aim to advance the field of XMORL. The following sections outline the work completed so far and the planned future directions.

4.1 Work Completed

Our work so far has focused on understanding foundations of the two underlying areas, MORL and XRL, and publishing a case-study paper in each.

Multi-Objective Deep Reinforcement Learning for Autonomous Systems: In my first publication [20], we applied Deep W-Learning (DWN) [8], a MORL technique, to a real-world autonomous system for the first time. Specifically, we integrated DWN into an Emergent Web Server (EWS) [3] self-adaptive exemplar, enabling runtime optimization of conflicting objectives such as response time and configuration cost. The results demonstrated that DWN can balance multiple objectives in complex environments, while avoiding issues associated with combining objectives into a single static utility function. This work highlighted both the potential of MORL for practical systems and the need to understand and trust learned policies.

Explaining Reinforcement Learning Decisions in Self-Adaptive Systems: I also contributed to the implementation of EARL (Explanations using Alternative Realities for Reinforcement Learning), a Python library designed to generate counterfactual explanations for RL-based self-adaptive systems. EARL enables the exploration of "what-if" scenarios to clarify agent behavior and decision-making in complex, realistic environments. The library implements multiple counterfactual generation methods, Variants of RACCER [4] and GANterfactual [9] and provides a model-agnostic interface to facilitate integration with RL agents. To demonstrate its applicability, we integrated EARL into Citi-bikes [10], a self-adaptive bike-sharing system simulation, providing explanations for agent decisions. This work has been submitted to a conference and is currently under review.

4.2 Future Work

The primary focus of my future research will be the development of explainability techniques specifically tailored for Multi-Objective Reinforcement Learning (XMORL). Building on the observations from our previous work, I will address research questions related to: (I) Designing explanations that capture both objective-specific reasoning and the trade-offs made by the agent, (II) Identifying specific application domains for XMORL, such as user-in-the-loop scenarios or settings requiring causal understanding, where XMORL techniques could be tested and would provide tangible benefits, (III) Adapting explanation methods to dynamic or evolving user preferences, (IV) Ensuring the scalability and effectiveness of explanations in high-dimensional objective spaces, and (V) Defining benchmarks for evaluating the quality of explanations in XMORL, that account for the unique challenges of multi-objective settings.

As an initial step, I am working on completing a position paper that formalizes the research gaps and challenges in XMORL where we identify possible intersections of both areas. My current work focuses on formalizing the research gaps that must be filled to address the problem of Explainable Multi-Objective Reinforcement Learning. My work aims to contribute towards the creation of trustworthy, explainable, and adaptive reinforcement learning agents capable of operating in complex, multi-objective real-world systems.

5 Conclusion

The intersection of Multi-Objective Reinforcement Learning and Explainable Reinforcement Learning presents both significant challenges and opportunities for the development of trustworthy autonomous systems. My work so far has explored this space by applying MORL techniques to real-world self-adaptive systems, contributing to the development of explanation tools for RL, and identifying open research areas in the emerging field of Explainable Multi-Objective Reinforcement Learning.

The results of these efforts highlight the potential of MORL for handling complex, multi-objective decision-making, and also express the potential limitations of existing explainability approaches when applied to such settings. As a result, the need for dedicated XMORL techniques that can make agent behavior more transparent, interpretable, and user-aligned is clear.

For the next steps of my research, I will focus on addressing these challenges by developing explanation methods tailored to MORL scenarios. These methods will aim to provide both objective-specific insights and an understanding of the trade-offs involved.

Acknowledgements

This publication has been supported in part by the Science Foundation Ireland under Grant number 18/CRT/6223 For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

[1] A. Abels, D. Roijers, T. Lenaerts, A. Nowé, and D. Steckelmacher. Dynamic weights in multi-objective deep reinforcement learning. In *International conference on machine learning*, pages 11–20. PMLR, 2019.

[2] R. M. Annasamy and K. Sycara. Towards better interpretability in deep q-networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4561–4569, 2019.

[3] R. R. Filho, E. Alberts, I. Gerostathopoulos, B. Porter, and F. M. Costa. Emergent web server: An exemplar to explore online learning in compositional self-adaptive systems. In *Proceedings of the 17th Symposium on Software Engineering for Adaptive and Self-Managing Systems*, pages 36–42, 2022.

[4] J. Gajcin and I. Dusparic. Raccer: Towards reachable and certain counterfactual explanations for reinforcement learning. *23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)*, 2023.

[5] J. Gajcin and I. Dusparic. Redefining counterfactual explanations for reinforcement learning: Overview, challenges and opportunities. *ACM Computing Surveys*, 56(9):1–33, 2024.

[6] X. He and C. Lv. Toward personalized decision making for autonomous vehicles: a constrained multi-objective reinforcement learning technique. *Transportation research part C: emerging technologies*, 156: 104352, 2023.

[7] T. Hickling, A. Zenati, N. Aouf, and P. Spencer. Explainability in Deep Reinforcement Learning: A Review into Current Methods and Applications. *ACM Computing Surveys*, 56(5):1–35, May 2024. ISSN 0360-0300, 1557-7341. doi: 10.1145/3623377.

[8] J. Hribar, L. Hackett, and I. Dusparic. Deep w-networks: Solving multi-objective optimisation problems with deep reinforcement learning. In *International Conference on Agents and Artificial Intelligence*, 2022.

[9] T. Huber, M. Demmler, S. Mertes, M. L. Olson, and E. André. Ganterfactual-rl: Understanding reinforcement learning agents' strategies through visual counterfactual explanations. *arXiv preprint arXiv:2302.12689*, 2023.

[10] A. Jiang, J. Zhang, P. Yu, L. Huang, Y. Qiu, J. Wang, W. Shi, K. Li, Z. Wang, C. Zhang, T. Sun, M. Chen, K. Yu, X. Wei, M. Li, N. Shang, Q. Meng, S. Li, J. Bian, B. Cheng, and T.-Y. Liu. Maro: A multi-agent resource optimization platform, 2020. URL https://github.com/microsoft/maro.

[11] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, and F. Doshi-Velez. Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on explainable artificial intelligence*, 2019.

[12] G. Liu, O. Schulte, W. Zhu, and Q. Li. Toward Interpretable Deep Reinforcement Learning with Linear Model U-Trees, July 2018. URL http://arxiv.org/abs/1807.05887. arXiv:1807.05887 [cs].

[13] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2493–2500, 2020.

[14] S. Milani, N. Topin, M. Veloso, and F. Fang. Explainable Reinforcement Learning: A Survey and Comparative Review. *ACM Computing Surveys*, 56(7):1–36, July 2024. ISSN 0360-0300, 1557-7341. doi: 10.1145/3616864.

[15] Y. Oh, A. Ullah, and W. Choi. Multi-objective reinforcement learning for power allocation in massive mimo networks: A solution to spectral and energy trade-offs. *IEEE Access*, 2023.

[16] X. Peng, M. O. Riedl, and P. Ammanabrolu. Inherently explainable reinforcement learning in natural language. *CoRR*, abs/2112.08907, 2021. URL https://arxiv.org/abs/2112.08907.

[17] E. Puiutta and E. M. Veith. Explainable Reinforcement Learning: A Survey, May 2020. URL http://arxiv.org/abs/2005.06247.

[18] M. Reymond, E. Bargiacchi, and A. Nowé. Pareto conditioned networks. *arXiv preprint arXiv:2204.05036*, 2022.

[19] M. Reymond, C. F. Hayes, L. Willem, R. Rădulescu, S. Abrams, D. M. Roijers, E. Howley, P. Mannion, N. Hens, A. Nowé, et al. Exploring the pareto front of multi-objective covid-19 mitigation policies using reinforcement learning. *Expert Systems with Applications*, 249:123686, 2024.

[20] J. C. Rosero, N. Cardozo, and I. Dusparic. Multi-objective deep reinforcement learning optimisation in autonomous systems. In *2024 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*, pages 97–102, 2024. doi: 10.1109/ACSOS-C63493.2024.00038.

[21] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

[22] C. van Rossum, C. Feinberg, A. A. Shumays, K. Baxter, and B. Bartha. A novel approach to curiosity and explainable reinforcement learning via interpretable sub-goals. *arXiv preprint arXiv:2104.06630*, 2021.

[23] Y. Wen, S. Li, R. Zuo, L. Yuan, H. Mao, and P. Liu. Skilltree: Explainable skill-based deep reinforcement learning for long-horizon control tasks. *arXiv preprint arXiv:2411.12173*, 2024.

[24] K. Zhang, J. Zhang, P.-D. Xu, T. Gao, and D. W. Gao. Explainable ai in deep reinforcement learning models for power system emergency control. *IEEE Transactions on Computational Social Systems*, 9(2): 419–427, 2022. doi: 10.1109/TCSS.2021.3096824.