Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD27834

Reinforcement Learning Interpretability Methods and Decision Making Methods under Constraints

Siheng Ye

Leeds College, Southwest Jiaotong University, Chengdu, China
map1e@my.swjtu.edu.cn

Abstract. Reinforcement learning (RL), as a core technology of artificial intelligence, has shown strong potential in the fields of robotics, games and autonomous driving. However, the "black box" nature of deep RL models leads to a lack of transparency in the decision-making process, making it difficult for users to understand and trust the agent behavior of RL models, and the uninterpretability of decisions may cause serious consequences in sensitive fields such as healthcare and finance. At the same time, because traditional RL pursues maximum reward and result models often ignore fairness, leading to policy bias, which affects the group's rights. So this article will summarize from the perspective of two key transparency and fairness of RL as summarized in the paper: one is based on the interpretability of the decision-making method, using the causal analysis and partial interpretation and visualization tools to make decisions transparent; Second, the decision-making method based on the constraint conditions, through multi-objective optimization and gradually constraints ensure the decision unfair. This review covers the methodologies, experimental results and limitations of representative literature in recent years. The significance of this paper is to systematically integrate these methods, reveal the interaction challenges of transparency and fairness, promote the development of more reliable RL systems, and look forward to future directions to help promote the ethical deployment and sustainable innovation of RL in social applications.

Keywords: Reinforcement learning, Interpretability, Decision making Introduction

1. Introduction

Reinforcement learning (RL) through the Markov decision process of interaction with the environment, to maximize the cumulative rewards discounted as the goal, has become an efficient sequential decision-making making in artificial intelligence learning. In recent years, the integration of deep neural networks has significantly improved the policy adaptation and generalization ability of RL in complex high-dimensional tasks. For example, in the scenarios where autonomous vehicles navigate complex road conditions and agents learn cooperation and confrontation in multi-player soccer games [1], RL has shown strong learning potential. However, deep RL policy networks often exhibit "black-box" characteristics, and their decision logic is difficult to intuitively understand by end users. Taking StarCraft II as an example, users often cannot explain why the agent chooses to "build a supply station" instead of "train combat units" at a particular moment, which weakens trust

DOI: 10.54254/2755-2721/2025.LD27834

and hinders human-machine collaboration [2]. This cannot be interpreted not only limiting the RL in key areas of application, but also magnifying the potential risks, such as strategy holes or fair.

The unexplainability of deep RL stems from three main aspects. First, the coupling of high-dimensional state-action space with deep network parameters leads to an elusive mapping path, and the agent may learn "shortcut" strategies due to reward shaping holes - e.g., in the MuJoCo Walker task, the agent deliberately falls and then gets up to obtain a higher reward shaping term instead of walking steadily [3]. Secondly, the cumulative effect of temporal decisions magnifies the difficulty of causal traceability. Experiments show that in the classical MountainCar environment, if the discount factor is not set properly, the agent will converge to the unexpected oscillation strategy, further increasing the explanation burden [4]. Finally, the problem of lack of fairness has become increasingly prominent. In the multi-agent cooperation scenario, when each agent independently optimizes its own payoff, the system is easy to fall into the "efficiency first" Nash equilibrium, resulting in resource allocation skewed to the majority group [5]. Especially in group-sensitive RL environments, if the environment transition kernel itself contains direct dependence on sensitive attributes (which violates the dynamic fairness condition), the long-term reward gap will continue to be magnified even if the strategy satisfies the immediate statistical fairness constraint [6].

To address the above challenges, this paper reviews the transparency and fairness research in RL from two complementary paths: first, from the perspective of interpretability. Based on causal reasoning of facts explain the importance [2], the time step to extract the [7,3] and partial strategies such as method, help users to establish a transparent DRL model; Secondly, the fairness constraint perspective. By the average reward of the multi-objective RL constraints [4], distributed social welfare maximization [5], and dynamic correction [6], fair optimization of cumulative returns at the same time ensures fairness. On this basis, this paper further explore the synergy potential of the two paths. Can be interpreted tools can provide intuitive basis for fair constraints, reveals how constraints affect every step of the decision; On the other hand, the introduction of fairness constraints can inject a new semantic dimension into the explanation mechanism, so that it can not only explain "why a certain group must be treated differently", but also clarify "why a certain group must be treated differently".

Furthermore, this paper analyze the existing limitations, including the computational bottleneck of counterfactual interpretation and the performance tradeoff brought by fairness constraints, and look forward to future directions: using large language models to automatically translate numerical explanations into natural language narratives, or designing adaptive fairness mechanisms in non-stationary environments. Comb through the system and method of fusion, this paper is intended to build a credible and impartial reinforcement learning system to provide the methodological framework and practical guide.

2. Interpretability based decision making method

In the field of reinforcement learning (RL), the development of interpretability methods aims to bridge the gap between model complexity and human understanding. These methods are typically categorized by explanation type, such as global explanations that provide holistic policy insight such as revealing temporal relationships through causal models, while local explanations profile specific decisions such as factorization based on feature contributions. They can also be divided into rear explanations according to the mechanism, such as training after the analysis of counterfactual simulation or displayed shapes value calculation, and the intrinsic interpretation, such as training embedded in the modular network design. The following review selected representative literature, focusing on the innovation of methodology, experimental verification and comparison to each other.

DOI: 10.54254/2755-2721/2025.LD27834

The introduction of the causal reasoning framework, the causal analysis method of lens will RL decision modeling for cause and effect diagram, by the fact that simulation generated explanation. It extracts causal relationships from the state-action-reward chain, and uses intervention actions to infer what the outcome would be if other actions were selected. This innovation is to combine causal models with the temporal properties of RL, avoiding the static limitations of traditional feature importance methods. In the training phase, the causal model needs to integrate the environment dynamics, while Monte Carlo sampling is used to approximate the causal path for explanation generation. This method is suitable for discrete action Spaces, such as Starcraft II. In the Starcraft II benchmark, the method evaluates task prediction accuracy, explanation satisfaction, and trust through a user study with 120 participants. The results show that the causal model significantly outperforms the baseline model in task prediction and satisfaction, but the impact on trust is not statistically significant. Compared with pure SHAP methods, this framework pays more attention to temporal causality and provides richer narrative explanations, but has higher computational overhead because multiple paths need to be simulated, which is not as efficient as local methods in real-time applications [2].

The local explanation framework is based on a variant of SHAP value, which provides feature contribution analysis for RL local decision making, and decomposes the value function into a weighted sum of state features. Its innovation lies in adapting to the sequential nature of RL, and calculating the impact of actions in a specific state through time-series sensitive SHAP. This framework uses a post-interpretation mechanism, uses background dataset sampling to calculate SHAP values, and supports continuous or discrete space. Its core formula is the expected difference of value contribution, which ensures the fidelity of interpretation. In the Four Rooms, Door-Key, MiniPacman, and Pong environments, the method evaluates the effectiveness of explanations through user studies, and the results show that it outperforms baselines in task prediction. This method performs well in low-dimensional tasks, but the sampling complexity increases in high-dimensional scenarios, leading to efficiency bottlenecks [7].

The EDGE framework focuses on the interpretation of the importance of time steps, and analyzes the activation patterns of deep RL networks by using a self-explanation model, which includes a Gaussian process and a custom kernel to generate explanations that highlight the influence of key time steps on the final reward. The innovation of EDGE is to bridge the Gaussian process and RL explanation, and transform abstract parameters into time-step level policy insights. This framework is integrated into the policy network, and parameter learning is optimized through variational inference and induced points. It is suitable for visual input tasks and supports end-to-end training. On Atari and MuJoCo environments, the method validates interpretation fidelity and demonstrates advantages in policy forensics through complementary user studies. Tests reveal that the method is robust in complex environments, but has weak generalization to non-sequential data [3].

The neural module pipeline method adopts a modular decomposition strategy to split the RL network into functional modules, such as exploration modules and reward evaluation modules, which provide functional explanations through induction and detection. Its innovation lies in emphasizing the transparency of the interaction between modules, which is borrowed from neuroscience. This pipeline includes eliculation, where regularized training promotes modularity; Detection, or activation analysis; And representation, which is function labeling. It uses decomposition techniques to quantify module contributions. In the MiniGrid environment, including 2D and 3D variants, the method reveals the emergence of navigation modules and validates functionality through ablation experiments. This method performs well in multi-task RL, but the training overhead increases [8].

DOI: 10.54254/2755-2721/2025.LD27834

The common advantage of these methods is to enhance the confidence of RL, but the limitations include computational cost and generalization challenges. In comparison, causal and local methods are more suitable for analytical tasks, while time-stepping and modularity are more suitable for engineering applications.

3. Decision making method based on constraints

The introduction of fairness constraints in reinforcement learning (RL) aims to alleviate the problem of bias amplification, where algorithmic decisions can amplify inherent biases in the data, leading to long-term inequality. These methods are classified by constraint type, such as hard constraints strictly enforce fairness, while soft constraints are achieved through regularization. They can also be classified into single-agent and multi-agent by application scenario. The following literature review highlights their optimization strategies, fairness metric improvements, and cross-comparisons. Analysis shows that although these methods balance fairness and performance, they often increase the difficulty of optimization. In the future, it is necessary to explore adaptive constraints to cope with dynamic environments.

Fairness policy learning in multi-objective RL regards fairness as multi-objective optimization, and uses average or discounted reward to constrain the Pareto front to solve. The innovation of multi-objective RL is to deal with the fairness effect of the discount factor and avoid the short-term bias of traditional RL. This approach embeds fairness regularities into the loss function and iterates through evolutionary algorithms or gradient descent. It is suitable for multi-objective scenarios, such as resource allocation. In the grid world, the fairness metric is significantly improved, which is based on the GGF welfare function with less reward loss. The results prove the effectiveness of the constraint, but the convergence of multi-objective search is slow [4].

The decentralized multi-agent fairness strategy adopts a decentralized framework and imposes fairness constraints through inter-agent communication to avoid a central bottleneck. Its innovation lies in distributed optimization, which improves global performance while maintaining local fairness. In this approach, a communication protocol and fair Lagrange multipliers are introduced, and agents update their strategies independently. It supports cooperative or competitive scenarios. In the navigation task, fairness improves significantly, the CV metric decreases, and the performance outperforms the baseline. This method is efficient in large-scale agents, but the communication overhead needs to be optimized, and experiments show that it may amplify latency in complex environments [5].

Dynamic fairness constrained RL explores dynamic fairness, captures the inequality in the environment dynamics, and evaluates the impact of changes in sensitive attributes on the next state and reward. Its innovation is to adopt a causal perspective, decompose the sources of inequality, distinguish the inequity in decision-making, historical and dynamic factors, and adapt to the temporal environment. This method introduces sensitive attribute intervention into the RL framework, and derives a recognition formula to reliably estimate from data. It is suitable for temporal tasks involving sensitive attributes, such as recommender systems. The bias is significantly reduced and the reward is stable. The results are real-time, but the dynamic robustness needs to be verified. Compared with the aforementioned global methods, this dynamic strategy is more flexible and easier to integrate causal analysis, but may amplify noise in multiple agents, which is complementary to decentralized multi-agent fair strategies [5,6]. The limitation is that it is difficult to adjust the constraint strength parameter, and an adaptive threshold is recommended.

The framework of the under-specification problem focuses on the hidden danger of fairness caused by the under-specification of the model, which is alleviated by diversified training. This

DOI: 10.54254/2755-2721/2025.LD27834

framework can be analogously applied to RL reward design, emphasizing the fairness of data and reward design. It analyzes distribution shifts and introduces regularized samples. In RL, it can be applied to reward reshaping. The bias is significantly reduced, providing theoretical support, and the results guide constraint design in RL evaluation [9].

These methods advance fair RL, but the performance tradeoffs such as increased computational complexity and complexity are generic issues, e.g., multi-objective optimization requires more iterations, and abstract frameworks are difficult to quantify directly. The comparison shows that multi-objective is suitable for a single environment, and decentralized or progressively more suitable for multi-agent environments.

4. Existing limitations and future perspective

Although existing approaches have made progress, they still face multiple limitations. Interpretability mechanisms often introduce additional computational overhead, such as causal graph construction or SHAP value calculation, which may lead to delays in real-time RL applications [10]. Although fairness constraints improve fairness, they may sacrifice the overall performance, especially in resource-limited environments, where optimization convergence slows down [11]. For example, this work proves exponential time lower bounds under exact fairness constraints, leading to significant performance degradation of learning algorithms in multi-state environments. In multi-agent scenarios, the interaction between explanations and constraints is complex, and it is difficult for a single method to satisfy both global transparency and local fairness [12]. The lack of adaptability in dynamic environments is also a pain point, and the stability of explanations is difficult to guarantee when agent behavior changes over time [13]. In addition, in continuous action Spaces, constraints may lead to policy instability; There are also conflicts between reward shaping and fairness, for example, historical data bias may be amplified by the reward function. From an ethical perspective, the paper may discuss fairness, but the interpretation method itself may introduce new biases (e.g., interpretation bias). For example, causal approaches assume that causal structure is known and may reinforce human biases. Literature coverage is biased towards methodologies and ignores application cases (e.g., fair RL in real-world deployments), which may lead to ignoring cross-group fairness issues [9]. For example, in autonomous driving, fairness RL needs to consider cross-fairness to avoid ethical dilemmas for vulnerable groups [12], and this work explains multi-agent behaviors through temporal queries, which can be extended to cross-group scenarios.

Looking forward to the future, the first is to integrate large language models (LLMS) to generate natural language explanations, such as combining GPT variants to describe RL decision logic to improve user friendliness. The second is to develop an adaptive constraint mechanism, which uses online learning to dynamically adjust the fairness threshold and adapt to environmental variation [11]. The third is to explore a hybrid framework for causal fairness that fuses transparent methods with constraints, such as using temporal causal analysis in multi-agent RL [12] to ensure long-term fairness, and borrowing from the taxonomy of cross-fairness [9] to extend to continuous attributes. At the same time, MDP transformation [13] can be fused with LLM to generate natural language explanations. In addition, the establishment of standardized evaluation benchmarks will facilitate method comparison and promote cross-domain applications, such as RL in environmental justice. These directions are expected to address current bottlenecks and enable the ethical transformation of RL.

DOI: 10.54254/2755-2721/2025.LD27834

5. Conclusions

The opacity and unfairness of reinforcement learning have become obstacles to its large-scale application. This paper reveals the root causes of these problems through background analysis, and gives a systematic review from two dimensions of interpretability and constraints. In terms of interpretability, methods such as causal explanation and local analysis effectively improve the decision transparency and help users understand the agent's logic deeply. In terms of constraints, multi-objective optimization and a dynamic fairness mechanism ensure the fairness of the strategy and reduce the influence of bias. The integration of these perspectives not only demonstrates the synergies between methods, such as transparent tools that reveal the side effects of constraints, but also highlights their practical implications: enhancing the trustworthiness of RL systems, complying with ethical standards, and providing reliable support for high-risk domains such as healthcare and finance.

Nevertheless, current research still faces challenges such as computational inefficiency, performance tradeoffs, and multi-agent complexity, which also provide opportunities for innovation. This survey constructs a comprehensive framework that recognizes these challenges and provides guidance for future research, such as fusing large language models for multimodal interpretation or developing dynamically fair algorithms to adapt to real-time scenarios, ultimately inspiring more innovations to drive reinforcement learning toward a more inclusive and transparent direction, promoting its sustainable integration and beneficial impact in society.

References

[1] Kurach, K. et al. (2020) Google Research Football: A Novel Reinforcement Learning Environment. Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4501-4510.
[2] Madumal, P., Miller, T., Sonenberg, L. and Vetere, F. (2020) Explainable Reinforcement Learning Through a Causal Lens. Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2493-2500.
[3] Guo, W., Wu, X., Khan, U. and Xing, X. (2021) EDGE: Explaining Deep Reinforcement Learning Policies. Advances in Neural Information Processing Systems.
[4] Siddique, U., Weng, P. and Zimmer, M. (2020) Learning Fair Policies in Multi-Objective (Deep) Reinforcement Learning with Average and Discounted Rewards. Proceedings of the 37th International Conference on Machine Learning.
[5] Zimmer, M., Glanois, C., Siddique, U. and Weng, P. (2021) Learning Fair Policies in Decentralized Cooperative Multi-Agent Reinforcement Learning. Proceedings of the 38th International Conference on Machine Learning.
[6] Deng, Z., Jiang, J., Long, G. and Zhang, C. (2024) What Hides behind Unfairness? Exploring Dynamics Fairness in Reinforcement Learning. Proceedings of the 33rd International Joint Conference on Artificial Intelligence, pp. 3908-3916.
[7] Luss, R., Dhurandhar, A. and Miao, L. (2023) Local Explanations for Reinforcement Learning. Proceedings of the AAAI Conference on Artificial Intelligence, pp. 9002-9010.
[8] Soligo, A., Ferraro, P. and Boyle, D. (2025) Inducing, Detecting and Characterising Neural Modules: A Pipeline for Functional Interpretability in Reinforcement Learning. Proceedings of the 42nd International Conference on Machine Learning.
[9] D'Amour, A. et al. (2022) Underspecification Presents Challenges for Credibility in Modern Machine Learning. Journal of Machine Learning Research, 23(226): 1-61.
[10] Gohar, U. and Cheng, L. (2023) A Survey on Intersectional Fairness in Machine Learning: Notions, Mitigation, and Challenges. Proceedings of the 32nd International Joint Conference on Artificial Intelligence, pp. 6619-6627.
[11] Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J. and Roth, A. (2017) Fairness in Reinforcement Learning. Proceedings of the 34th International Conference on Machine Learning, pp. 1617-1626.
[12] Boggess, K., Kraus, S. and Feng, L. (2023) Explainable Multi-Agent Reinforcement Learning for Temporal Queries. Proceedings of the International Joint Conference on Artificial Intelligence, pp. 55-63.
[13] Finkelstein, M., Liu, L., Levy Schlot, N., Kolumbus, Y., Parkes, D.C., Rosenschein, J.S. and Keren, S. (2022) Explainable Reinforcement Learning via Model Transforms. Advances in Neural Information Processing Systems.