A Study of Explainability Inquiry Based on Reinforcement Learning

Shuqi Yang

Leicester International Institute, Dalian University of Technology, Panjin, 124221, China

Abstract. With the application of Reinforcement Learning (RL) in security-sensitive fields such as healthcare, autonomous driving, and finance, the lack of Explainable Reinforcement Learning (XRL) restricts the technology's grounding and social trust. This paper reviews XRL's progress, constructing a framework covering core challenges, methods, and applications. Core challenges include black-box decision-making, reward bias, and complex multi-intelligence interaction. Among the existing methods, the intrinsic interpretability of the model is limited by the generalization ability, the ex-post interpretation method faces the problem of poor local-global consistency, and the hybrid method is difficult to dynamically adapt due to the high complexity of the system. XRL shows potential in medical decision-making, autonomous driving safety, and financial risk control through causal reasoning and multimodal interpretation, but further optimization is needed. This paper emphasizes building a standardized evaluation system and explores cutting-edge directions like cross-domain migration and multimodal human-computer collaborative interpretation to deepen XRL's theoretical framework and promote its safe application in high-risk domains.

1 Introduction

RL is a general algorithm for solving environmental interaction problems by optimizing strategies through interaction with the environment [1]. However, its black-box modeling nature leads to a lack of transparency in the decision-making process, which raises multiple challenges such as policy reliability [2]. The absence of explainability hinders the practical implementation of RL technology and acts as an impediment to its large-scale application in critical areas. There is an urgent need for systematic research to bridge the gap between technical performance and human cognitive requirements.

Recent years have seen the emergence of research in XRL, however, the methodological system remains underdeveloped. Liu Xiao et al [1] systematically analyzed the fundamental problems of XRL, pointed out that the temporal dependence of black-box decision-making and the bias of reward mechanism are the core challenges, and proposed a two-level classification framework of model intrinsic interpretability and post hoc interpretation methods. Tang Lei et al [3] constructed a classification system of RL interpretable methods through systematic discussion, which unfolded in terms of interpretable models, policy-

3422539153@mail.dlut.edu.cn

value functions, and environment interactions, and emphasized the potential of environment interaction-based interpretations in the understanding of intelligences' behaviors, but the interpretation of global states in dynamic and complex environments still needs further refinement. At the application level, Komorowski et al. [4] proposed a deep reinforcement learning-based framework (AI Clinician), which generates treatment strategies through the integration of Q-learning and neural networks. Whilst extant research has advanced the development of XRL from diverse perspectives, there is a paucity of holistic syntheses of the challenges, methodologies and applications of the technology. Moreover, the fragmentation of evaluation criteria poses a significant obstacle to technological transfer.

The purpose of this paper is to systematically review the research progress of XRL, and construct an analytical framework from the core challenges, method classification, application scenarios, and frontier directions. Firstly, an analysis of the core challenges of XRL is conducted, including black-box decision-making, reward bias, and multi-intelligence complexity. Secondly, a summary of the methods and limitations of intrinsic, post hoc, and hybrid explanations of the model is provided. Thirdly, a summary of the scenarios of typical high-security and sensitive domains, such as healthcare and autonomous driving, is presented. Finally, a discussion of the cutting-edge direction of causal Reinforcement Learning and multimodal explanations is offered. This paper provides a reference for deepening XRL research and promoting its safe deployment in high-risk domains.

2 Core Challenges in Explainable Reinforcement Learning

The core challenges of XRL are as follows: decision complexity, reward bias, and multi-intelligence body collaboration. These issues are particularly salient in domains characterized by heightened security sensitivity, exerting a direct influence on the technical credibility and practical value of the research. In this paper, we analyze three aspects of decision-making: the black box, reward bias, and multi-intelligence body interaction.

The intricacy of black-box decision-making functions as the principal impediment to the implementation of XRL. The Deep Reinforcement Learning model utilizes neural networks for high-dimensional feature extraction and policy optimization. The internal parameter interactions and timing-dependent characteristics of the model result in a decision logic that is difficult to trace [1]. To illustrate this point, consider the proximal policy optimization (PPO) algorithm. The nonlinear mapping of the policy gradient update complicates the intelligent body's behavior, making it difficult to express in explicit, logical terms, and, as a result, not easily understandable by users [2]. Furthermore, the temporal decision-making nature of RL exacerbates the difficulty of causal attribution. Intelligentsia decision-making while performing sequential tasks is influenced by the accumulation of historical states and actions, leading to difficulty in separating the causality of decision paths [3].

Implicit bias in the reward mechanism is the central causative factor for RL strategies to deviate from the real needs. The design of the reward function, as a key mechanism to guide the behavior of an intelligent body, is often biased due to the mismatch with human intentions [3]. In the context of financial trading strategies, when the objective of the reward function is exclusively focused on maximizing returns, an intelligent body may attain short-term gains through high-frequency arbitrage strategies. However, this approach may overlook crucial considerations such as market liquidity and risk management [5,6]. This discrepancy can be attributed to the mismatch between the designed reward function and the actual reward function. The complexity of environmental constraints, which are difficult for designers to fully capture, often fails strategies in practical application scenarios [7].

Multi-agent interaction and coordination represent a significant bottleneck hindering the large-scale deployment of reinforcement learning. In Multi-Agent Reinforcement Learning (MARL) scenarios, unpredictable group strategies emerge from the collaborative and competitive behaviors among intelligences [8]. In the context of vehicle-road cooperative systems, the effective exchange of dynamic information in real time among vehicles is imperative for the optimization of global traffic. However, discrepancies in local observations, as well as conflicting interests, may result in divergent interpretations of logical principles [9]. Furthermore, the strategic interplay between intelligences complicates the interpretation of a single behavior, hindering its ability to reflect the comprehensive global impact. This complexity is particularly evident in scenarios such as financial high-frequency trading. In the context of concurrently operating multiple trading algorithms, the local optimality of individual strategies can potentially precipitate systemic risk [5].

3 A Framework for the Categorization of Interpretable Methods

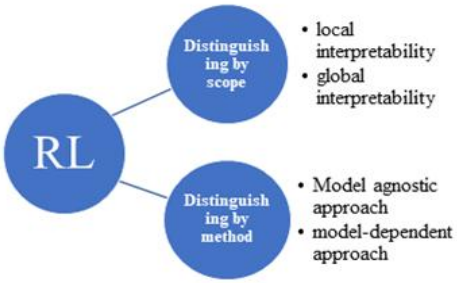

Fig.1 Common RL Classifications(Photo credit : Original)

The objective of this section is to systematize the current research status of XRL methods and to classify XRL methods into three categories: model intrinsic interpretability, ex post interpretation methods, and hybrid interpretation systems. These categories are based on the three dimensions of generation timing, model dependency, and interpretation granularity. The present classification of XRL methods is predicated on two fundamental issues. First, the prevailing classification system predominantly emphasizes a solitary technical trajectory, such as the categorization by model structure or explanation generation, as illustrated in Figure 1. This approach exhibits an absence of comprehensive integration of method interactions and application scenarios. Second, the fragmentation of evaluation criteria leads to the difficulty of cross-domain method migration. In this section, a unified classification framework is constructed to clarify the principal boundaries, advantages, limitations, and application scenarios of various methods. This methodological support will provide a foundation for the subsequent analysis of application scenarios.

Model Intrinsic Interpretability is a model-dependent approach that makes the decision-making process of intelligent systems directly understandable to humans. It also makes the decision logic interpretable during the training process by constructing models with intrinsically interpretable structures. Current research focuses on two technical approaches: strategy distillation methods and micro-white-box models. Strategy distillation methods

convert black-box strategies into interpretable models via knowledge distillation. Typical examples include decision tree distillation and linear model compression. Cao, Hongye, et al. noted that the VIPER algorithm uses an imitation learning framework to extract policy rules from deep Q-networks and build a decision tree model. This model can essentially maintain the original policy's performance by explicitly revealing the state-action mapping relationship through feature importance ranking in the CartPole and Lunar Lander tasks [2]. On the other hand, micro-white-box models employ a transparent architecture that can be gradient-optimized, facilitating end-to-end policy learning. A notable example is Differential Decision Trees (DDTs), proposed by Bekkemoen, which utilize microprocessible nodes to replace the hard-threshold splitting mechanism of traditional decision trees. Their efficacy is evidenced by significant performance enhancement over CART decision trees in continuous control tasks, such as MuJoCo Ant [10]. However, the prevailing approach continues to confront substantial challenges, including the salient performance-interpretation trade-off predicament. The performance of the white-box model is constrained by its structural simplification, resulting in lower performance compared to the black-box model. Additionally, the model's capacity to generalize and adapt to dynamic environments remains limited [10]. Future research endeavors must explore hierarchical explanation frameworks and online rule evolution mechanisms to achieve a balance between explanation granularity and policy effectiveness in complex tasks [2,10].

Post-hoc interpretation methods provide complementary interpretations by analyzing the behavior of trained models. These methods focus on the reverse parsing of trained models and have the advantage of being compatible with all types of black-box models without modifying the original model architecture. These methods can be divided into two categories according to their level of detail: local and global. Local methods characterize the dependencies of specific decisions by altering the input features. A systematic classification study by Tang Lei et al. showed that an improved RL-based explanation framework (e.g., LIME-RL) can successfully identify the sensitivity of state features to Q-value changes in the GridWorld navigation task [3]. This is achieved by constructing a local linear agent model, which promotes the application of XRL methods to robot path planning. It also validates decision-making in safety-critical scenarios. A theoretical tool is provided. Global explanatory methods reveal the model's overall behavioral pattern at the macro level through visualization techniques (e.g., attention heat map, Monte Carlo tree search path). Cao, Hongye, et al. reviewed an attention heat map–based approach that uses visualization to diagnose faults in robotic arm control systems by identifying critical joint motion trajectories in a MuJoCo robotic arm control task [2]. However, the a posteriori method has several drawbacks, including low interpretation fidelity, interpretation fidelity dependent on sampling accuracy, poor consistency between local and global interpretations, and high computational complexity in high-dimensional state spaces [1].

Hybrid interpretation systems achieve a balance between real-time performance and the depth of interpretation by integrating model intrinsic interpretability with post hoc interpretation methods. Among these techniques, neural-symbolic integration is a recently developed hybrid interpretation technique that generates interpretations that are easily comprehensible while maintaining high model performance. This is achieved by combining the powerful representation capability of neural networks and the logic of symbolic reasoning. The concept of neuro-symbolic integration has been proposed as a potential solution for explainable reinforcement learning. This integration involves the combination of the advanced representation capabilities of neural networks with the logic of symbolic reasoning. In domains such as autonomous driving, medical decision-making, and multi-intelligent body systems, neural symbolic approaches have been shown to enhance the transparency of models, thereby improving their adaptability and robustness. However, the

integration of these two methods in complex dynamic environments remains a challenge for current research [9,11].

Table 1 Comparison of XRL Methods and Applicable Scenarios

classifications	Timing of generation	Model dependency	Explain granularity	Advantages	Limitations	Typical application scenarios
Model Intrinsic Interpretability	Training phase	Strong dependence	global interpretation	High interpretation fidelity and real-time	Limited model performance, poor generalization to dynamic environments	Medical decision making, robotic arm control
Post-hoc interpretation methods	After training	Model-independent	Global/partial explanation	Compatible with black-box modeling	Localized interpretations may mislead global understanding	Financial Trading Strategies
Hybrid interpretation systems	Training and inference phase	Partial model dependency	Dynamic Multi-Granularity	Compatible with real-time and deep analytics	High system complexity	Multimodal Robot Collaboration Scenario

Table 1 presents a comparative analysis of the characteristics and application scenarios of the three types of methods. As demonstrated in the table, model intrinsic interpretation methods demonstrate exceptional proficiency in terms of interpretation fidelity and real-time performance. However, these methods are constrained by the limitations of the model's performance and generalization capacity. Conversely, ex post interpretation methods exhibit compatibility with black-box models and exhibit a high degree of flexibility. Nevertheless, these methods encounter the challenge of maintaining consistency between local and global interpretations. The hybrid interpretation system possesses the capacity for dynamic, multi-granularity interpretations. However, the system's inherent complexity hinders its practical implementation.

4 RL in key areas

The potential for wide application of XRL in high-risk fields, such as healthcare, autonomous driving, and finance, has been demonstrated. The methodological innovation of XRL is deeply coupled with the needs of these fields. This section illustrates the core role and methodological breakthroughs of XRL in various fields through typical case studies, and reveals the key challenges in the implementation of the technology.

In the domain of healthcare, XRL aspires to enhance transparency and causal explanation of treatment strategies, thereby promoting the acceptance of clinical applications. RL has the capacity to enhance the efficacy of treatment strategies and risk prediction models, thereby facilitating more precise medical decision-making. However, the black-box nature of the model imposes limitations on clinicians' comprehension of its decision logic, a situation that may give rise to ethical and trust crises. In an effort to address these limitations, Komorowski et al. (2018) have proposed a deep reinforcement

learning-based framework (AI Clinician) for generating sepsis treatment strategies through the integration of Q-learning with neural networks. The model utilizes dynamic clinical variables (e.g., blood pressure, infusion volume) from historical electronic health records (MIMIC-III) to train the strategy and optimize long-term survival. Clinical validation demonstrated that the decision pathway was highly compatible with the physician's cognitive model and significantly improved the trustworthiness of the treatment plan [4]. Cao Hongye et al. have indicated that XRL applied to medical scenarios must offer multimodal interpretations (e.g., visual decision paths versus natural language descriptions) to satisfy physicians' need for transparency in decision logic [2]. In complex clinical tasks, such as sepsis, XRL frameworks have the potential to enhance the efficiency of human-computer collaboration by reducing the opaque nature of strategies. However, further exploration is necessary to ascertain their clinical utility through real-world scenario validation and causal interpretability analysis. At the evaluation level, medical XRL must genuinely reflect the decision logic and be clinically compatible to ensure the fidelity and clinical utility of the explanations by clinical specifications [4,12].

The demand for XRL in the domain of autonomous driving is centered on safety verification and real-time transparency. RL has been demonstrated to offer advantages in terms of path planning and dynamic obstacle avoidance; however, the black-box feature may pose potential safety hazards. Song et al. (2023) have proposed an interpretable autonomous driving framework based on deep reinforcement learning（DRL）. In a high-speed driving simulation environment, the framework has been demonstrated to learn to generate logic rules based on information such as the position and speed of surrounding vehicles in a lane change scenario. This suggests that the framework is capable of good decision-making performance and provides transparency and interpretability guarantees for the decision-making process of autonomous driving [11]. The extraction and validation of such logic rules provide a clear decision path for the automatic driving system, ensuring its compliance with preset safety thresholds and behavioral norms, and enhancing the transparency, safety, and reliability of the system.

The necessity for XRL in the financial sector is driven by two fundamental imperatives: regulatory compliance and risk management. RL demonstrates proficiency in high-frequency trading and asset allocation; however, concerns have been raised regarding the potential implications of its strategies, namely the possibility of market manipulation. Xu Bo et al. initiated their study by examining a variety of traditional DRL algorithmic models. These models were employed to address high-dimensional data in financial markets, accurately capture market dynamics characteristics, and dynamically adjust trading decisions. In the domain of quantitative trading in financial markets, DRL models have been shown to enhance the performance and return on investment of trading strategies. These models also enhance, to a certain extent, the adaptability to market volatility and the accuracy of trading decisions. This, in turn, promotes the innovation and development of quantitative trading strategies in financial markets [13]. The XRL model explains trading behaviors through a causal response function. This function helps to quantify the impact of trading orders on the volatility of market prices. By explaining market fluctuations, traders can better understand the decision logic of RL models. As a result, they can optimize trading strategies [12].

Table 2 Drivers of XRL Demand and Common Assessment Criteria

Domains	Type of requirements	Typical cases	Portfolio of common assessment criteria
Medical decision-making	Clinical Traceability and Ethical Compliance	ICU treatment decisions	Causal interpretability, knowledge graph

			constraint validation
Automatic driving	Secure authentication and real-time transparency	Vehicle-Road Collaboration Testing	Formal security verification (temporal logic), cognitive alignment
Financial transactions	Regulatory Compliance and Risk Control	Explanation of high-frequency trading strategies	Market mechanism compatibility (liquidity impact), regulatory semantic mapping

Table 2 provides a synopsis of the categories of XRL requirements, exemplary scenarios, and combinations of evaluation criteria in the domains of healthcare, autonomous driving, and finance. For instance, the medical domain must ensure clinical traceability through causal interpretability and knowledge graph validation, while the financial domain must combine market mechanism compatibility and regulatory semantic mapping to achieve compliance. This table reveals the strong coupling of domain requirements and assessment criteria, providing a guiding framework for scenario adaptation of XRL technology.

5 Directions and open issues

The field of Artificial Intelligence has identified XRL as a key research direction, and it is anticipated that this area will continue to evolve and yield numerous open problems in the future.

Subsequent research endeavors may involve the exploration of the hierarchical explanation framework and the online rule evolution mechanism. These explorations could potentially address the limitations associated with the trade-off between performance and interpretability, thereby enhancing the applicability and effectiveness of XRL methods in a range of tasks. Concurrently, the absence of cross-domain migration functionality constitutes a significant challenge in contemporary XRL research. The majority of these methods are designed for a single scenario and lack a universal interface design, which complicates the direct migration of dynamic interpretation generation techniques to new domains, such as multi-intelligence collaboration scenarios [1,10]. The construction of a task-driven adaptive classification framework, the establishment of cross-domain evaluation benchmarks, and the realization of efficient knowledge migration and sharing are of great significance to promote the rapid development of XRL in the application field.

Furthermore, the application of causal reinforcement learning and multimodal explanation generation techniques represents a state-of-the-art approach within the domain of XRL. Causal Reinforcement Learning is expected to provide a more profound and reliable explanation basis for the decision-making of intelligences by exploring causal relationships, which is potentially valuable for solving the reward bias in RL and the problem of MARL interaction. Multimodal explanation generation technology integrates information from multiple modalities, including text, image, speech, and others. This integration contributes to enhancing the efficiency and reliability of human-computer collaboration. However, the development of this technology necessitates theoretical breakthroughs and technological innovations.

The development of a standardized evaluation system is imperative. Presently, the quantitative correlation between human evaluation and algorithm evaluation remains to be established, resulting in an absence of an objective basis for method preference [12]. In the future, efforts should be made to build a unified evaluation standard, organically integrate multi-dimensional indicators, provide a scientific and objective quantitative basis for the

performance evaluation and comparison of XRL methods, promote the synergistic development of XRL theory and practice, and promote its safe and reliable application in high-risk fields.

6 Conclusion

This paper undertakes a systematic review of the extant research on XRL. It also constructs an analytical framework that covers the following aspects: the core challenges, method classification, application scenarios, and evaluation criteria. A close examination of the literature reveals that the core challenges of XRL are centered on the complexity of black-box decision-making, reward mechanism bias, and the complexity of multi-intelligence interaction. Among the extant methods, the interpretability of models is constrained by the trade-off between performance and interpretation. The fidelity of ex-post interpretation methods is contingent on sampling accuracy, and hybrid methods are challenging to dynamically adapt due to the high complexity of the system. Although XRL has demonstrated potential for application in high-risk domains such as healthcare and automated driving, it still requires optimization by combining causal inference and multimodal techniques for clinical traceability, safety validation, and regulatory compliance.

Future research must overcome the performance-interpretation trade-off bottleneck. It is imperative to construct a standardized evaluation system that quantifies the correlation between human cognitive alignment and algorithmic fidelity and supports cross-domain method migration. Furthermore, the exploration of the neural symbol integration framework and the multi-intelligence body collaborative interpretation mechanism is anticipated to unveil novel methodologies for transparent decision-making in dynamic environments. This study provides a systematic reference for deepening XRL theory and lays a methodological foundation for promoting safe deployment in high-risk domains. This is of great academic value and practical significance for promoting the development of trusted AI.

References

1. Liu, X., Liu, S., Zhuang, S. K., et al.: 'Exploration of Interpretability Foundations for Reinforcement Learning and Review of Methods', *Journal of Software*, 2023, 34, (5), pp. 2300–2316. DOI: 10.13328/j.cnki.jos.006485
2. Cao, H. Y., Liu, X., Dong, S. K., et al.: 'A review of interpretability studies for reinforcement learning', *Journal of Computing*, 2024, 47, (8), pp. 1853–1882
3. Tang, L., Niu, Y. Y., Wang, R. J., et al.: 'Classification study of interpretable methods for reinforcement learning', *Computer Application Research*, 2024, 41, (6), pp. 1601–1609. DOI: 10.19734/j.issn.1001-3695.2023.09.0430
4. Komorowski, M., Celi, L. A., Badawi, O., et al.: 'The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care', *Nature Medicine*, 2018, 24, pp. 1716–1720. https://doi.org/10.1038/s41591-018-0213-5
5. Van Hasselt, H., Guez, A., Silver, D.: 'Deep reinforcement learning with double Q-learning', *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016, 30, (1). https://doi.org/10.1609/aaai.v30i1.10295
6. Ruggeri, F., Russo, A., Inam, R., Johansson, K. H.: 'Explainable reinforcement learning via temporal policy decomposition', *arXiv preprint*, arXiv:2501.03902, 2025

7. Song, L., Li, D. Z., Xu, X.: 'A review of inverse reinforcement learning algorithms, theories and applications', *Journal of Automation*, 2024, 50, (9), pp. 1704–1723. DOI: 10.16383/j.aas.c230081
8. Quan, J., Zhang, S., Hu, W., et al.: 'Intelligent flight control test of unmanned aircraft based on deep reinforcement learning', *Advances in Aeronautical Engineering*, 2025, pp. 1–14. http://kns.cnki.net/kcms/detail/61.1479.V.20250324.1414.002.html
9. Yuan, L., Zhang, Z. Q., Li, L. H., et al.: 'Advances in collaborative multi-intelligent body reinforcement learning in open environments', *Chinese Science: Information Science*, 2025, 55, (2), pp. 217–268
10. Bekkemoen, Y.: 'Explainable reinforcement learning (XRL): a systematic literature review and taxonomy', *Machine Learning*, 2024, 113, pp. 355–441. https://doi.org/10.1007/s10994-023-06479-7
11. Song, Z., Jiang, Y., Zhang, J., et al.: 'An Interpretable Deep Reinforcement Learning Approach to Autonomous Driving', *IEEE Transactions on Intelligent Vehicles*, 2023, 9, (3), pp. 456–468. https://doi.org/10.1109/TIV.2023.1234567
12. Kim, B.: 'Towards a rigorous science of interpretable machine learning', *arXiv preprint*, arXiv:1702.08608, 2017. https://doi.org/10.48550/arXiv.1702.08608
13. Xu, B., He, Y. J., Wen, J. C., et al.: 'A review of deep reinforcement learning applied to quantitative trading in financial markets', *Journal of Intelligent Science and Technology*, 2024, 6, (4), pp. 416–428