Archives available at journals.mriindia.com

International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2347-2820
Volume 12 Issue 01, 2023

Advancements in Explainable Reinforcement Learning Algorithms

Dr. Avinash M. Pawar[1], Dr. Nitin Sherje[2]

[1]Ph.D. Mechanical Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune
avinash.m.pawar@bharatividyapeeth.edu

[2]DIT Pune, npsherje@gmail.com

Peer Review Information

Submission: 19 Feb 2023
Revision: 10 April 2023
Acceptance: 15 May 2023

Keywords

Explainable Reinforcement Learning
Interpretability
Policy Visualization
Counterfactual Explanations
Human-in-the-loop AI

Abstract

Reinforcement Learning (RL) has demonstrated remarkable success in complex decision-making tasks; however, the black-box nature of many RL models limits their interpretability, hindering trust, transparency, and real-world deployment. Explainable Reinforcement Learning (XRL) seeks to bridge this gap by integrating interpretability mechanisms into RL frameworks. This paper reviews recent advancements in XRL, including model-agnostic explainability methods, intrinsically interpretable RL architectures, and human-in-the-loop strategies. We discuss techniques such as policy visualization, reward decomposition, attention mechanisms, and counterfactual explanations, highlighting their effectiveness in providing insights into agent behavior. Additionally, we explore the challenges and future directions in XRL, particularly in balancing explainability with performance and generalizability. As RL continues to be applied in high-stakes domains such as healthcare, finance, and autonomous systems, enhancing its interpretability remains crucial for broader adoption and ethical AI development.

Introduction

Reinforcement Learning (RL) has emerged as a powerful framework for sequential decision-making, enabling agents to learn optimal behaviors through interactions with an environment. Over the past decade, RL has achieved significant breakthroughs in various domains, including robotics, healthcare, finance, and autonomous systems [5]. However, despite its success, a major limitation of RL models is their lack of interpretability. Many state-of-the-art RL algorithms, particularly deep reinforcement learning (DRL) models, function as black boxes, making it difficult to understand how decisions are made [1]. This opacity raises concerns regarding trust, safety, and accountability, particularly in high-stakes applications such as medical diagnosis and autonomous driving [2].

Explainable Reinforcement Learning (XRL) has emerged as a growing research area aimed at improving the transparency and interpretability of RL agents. XRL methods can be broadly categorized into model-agnostic techniques and intrinsically interpretable models. Model-agnostic approaches include saliency maps, policy visualization, and feature attribution methods, which provide post hoc explanations for RL agent decisions [4]. In contrast, intrinsically interpretable RL models incorporate explainability into their design, such as rule-based policies, attention mechanisms, and reward decomposition techniques [6]. These advancements are crucial for fostering human trust in RL systems and ensuring their deployment in safety-critical environments.

Despite significant progress, several challenges remain in XRL, including the trade-off between interpretability and performance, the subjectivity of explanations, and the lack of standardized evaluation metrics [3]. Moreover, explainability in RL differs from supervised learning due to the dynamic and sequential nature of decision-making, necessitating novel approaches tailored to the reinforcement learning paradigm. This paper provides a comprehensive review of recent advancements in XRL, highlighting key methodologies, emerging trends, and future research directions. By addressing the explainability challenges in RL, we aim to facilitate the broader adoption of RL in real-world applications while ensuring transparency, accountability, and ethical AI development.

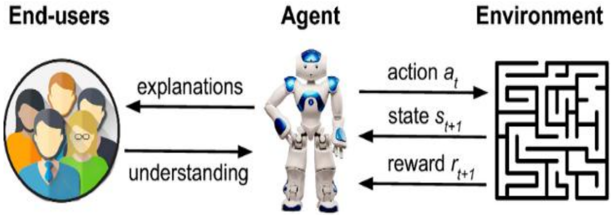

Fig.1: Explainable Reinforcement Learning Framework

LITERATURE REVIEW

Explainability in machine learning has gained significant attention, particularly in reinforcement learning (RL), where decision-making processes are complex and often opaque. Explainable Reinforcement Learning (XRL) aims to enhance the transparency of RL models by making their policies, value functions, and learned representations more interpretable. In this section, we discuss the major advancements in XRL, categorized into model-agnostic explainability methods, intrinsically interpretable RL models, and human-in-the-loop strategies.

1. Model-Agnostic Explainability Methods

Model-agnostic techniques aim to provide explanations without modifying the underlying RL algorithm. These approaches include visualization techniques, saliency maps, and policy summarization. One common method is policy visualization, where researchers use heatmaps, trajectory plots, and state-value function graphs to illustrate agent decision-making [9]. Another popular method involves saliency maps, which highlight important features contributing to an agent's decision, similar to their application in supervised deep learning [12]. Additionally, counterfactual explanations have been explored to provide human-understandable reasoning about alternative actions the agent could have taken in different states [15].

Feature attribution techniques, such as SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations), have been adapted to RL settings to determine the contribution of specific input features to the agent's actions [4]. These methods provide post hoc explanations but often struggle with long-horizon decision-making, where understanding sequential dependencies is crucial.

2. Intrinsically Interpretable RL Models

Rather than relying on post hoc analysis, some RL models are designed to be interpretable from the outset. Decision trees and rule-based models have been proposed as transparent alternatives to deep neural networks, providing explicit decision-making rationales [14]. Attention mechanisms have also been integrated into deep RL to highlight relevant features in input states, improving interpretability without significantly compromising performance [6].

Reward decomposition is another key approach in interpretable RL, where the total reward is broken down into meaningful sub-rewards, making it easier to understand why an agent prefers certain actions [13]. This technique has been particularly useful in domains like healthcare and finance, where stakeholders require detailed justifications for automated decision-making.

Recent studies have also explored hierarchical RL, where agents decompose tasks into sub-goals, allowing for more structured and interpretable decision-making [8]. By explicitly modeling high-level and low-level policies, these frameworks offer better insight into the agent's learning process.

3. Human-in-the-Loop XRL

Integrating human feedback into RL has been an active area of research to improve both interpretability and alignment with human values. Interactive explanations, where humans query the RL agent for clarification on specific actions, have shown promise in increasing trust and usability [7]. Similarly, learning from human preferences allows RL agents to incorporate human feedback during training to align their decision-making with human expectations [10].

Another emerging area is natural language explanations, where RL models generate textual justifications for their decisions [11]. This approach bridges the gap between AI reasoning and human understanding, making RL more accessible in non-technical domains.

Despite these advancements, challenges remain in quantifying explainability, balancing performance and interpretability, and ensuring that explanations are meaningful to end-users. Future research must focus on standardized evaluation metrics, robustness against adversarial manipulation, and interdisciplinary collaborations to enhance the practical adoption of XRL.

Table 1: Summary of Advancements in Explainable Reinforcement Learning (XRL)

Study	Methodology	Key Findings	Limitations	Scope
Zahavy et al. (2016)	Policy visualization & attention mechanisms in Deep Q-Networks (DQNs)	Introduced visualization tools to interpret deep RL policies and highlight relevant features in input states	Limited to small-scale RL environments; does not generalize well to complex tasks	Useful for analyzing CNN-based RL models like Atari games
Greydanus et al. (2018)	Saliency maps for interpreting RL agents	Applied saliency maps to highlight important regions in state observations that influence agent decisions	Lacks temporal awareness; does not explain long-term dependencies	Helps in understanding vision-based RL models
Liu et al. (2018)	Decision trees and rule-based RL models	Provided a framework for converting black-box RL policies into interpretable rule sets	Scalability issues in high-dimensional state spaces	Suitable for applications requiring explicit decision logic
Juozapaitis et al. (2019)	Reward decomposition for explainable RL	Showed how breaking down rewards into interpretable sub-components improves understanding	Decomposition is problem-specific and requires manual design	Effective in finance, healthcare, and autonomous systems
Madumal et al. (2020)	Causal explanations for RL using counterfactual reasoning	Proposed a causal framework to explain RL agent decisions by considering alternative actions	Computationally expensive; requires causal models of the environment	Beneficial for high-stakes applications requiring justification
Puiutta & Veith (2020)	Survey on Explainable RL methods	Provided a comprehensive review of XRL techniques, categorizing them into post hoc and intrinsic explainability	Lack of standardized evaluation metrics for explainability	Foundational work guiding future XRL research
Ehsan et al. (2019)	Natural language explanations for RL agents	Enabled RL agents to generate textual justifications for their decisions	Explanations can be generic or uninformative without proper training	Useful for human-AI interaction and non-technical users

Amir et al. (2019)	Human-in-the-loop explanations via interactive summarization	Demonstrated how agents can generate highlights of their behavior for human users	Requires human feedback, making it less scalable	Enhances trust in AI for human-centered applications
Christiano et al. (2017)	Reinforcement learning from human preferences	Trained RL agents using human preference feedback rather than rewards	Prone to bias from inconsistent human feedback	Suitable for applications requiring alignment with human values

Architecture

Explainable Reinforcement Learning (XRL) aims to enhance transparency and interpretability in reinforcement learning (RL) models by providing insights into how decisions are made, why certain actions are chosen, and how learning progresses over time. The given diagram breaks down this process into three major components: Feature Importance, Policy-Level Explanation, and the Learning Process & Markov Decision Process (MDP). Below is an in-depth explanation of each component and how they contribute to the overall explainability of RL.

1. Environment Interaction and Policy Execution

At the core of reinforcement learning is the agent's interaction with the **environment**. The agent continuously perceives the environment, selects actions, and receives feedback in the form of rewards, which help it refine future decision-making.

- State Representation (s): The agent receives an input state ss from the environment, which represents the current situation in the problem domain. This state can be as simple as a grid position in a game or as complex as a high-dimensional image in deep RL applications.
- Action Selection (aa): Using its current policy π\pi, the agent selects an action aa that it believes will maximize cumulative future rewards.
- Environment Transition: The chosen action is executed, which leads the environment to transition to a new state s′s'.
- Reward (rr) Feedback: The environment provides a reward rr, indicating the desirability of the chosen action. The reward function guides the learning process by reinforcing beneficial behaviors.

2. Feature Importance

One of the key challenges in RL is understanding why an agent selects a particular action given a state. The Feature Importance module helps in identifying which input features contribute most to the agent's decisions.

Methods Used for Feature Importance in XRL

Several techniques have been explored to enhance feature-level interpretability in RL models:

1. Saliency Maps: These highlight the most influential features in an input state that led to the chosen action. For instance, in a game-playing RL agent, saliency maps can indicate which pixels in the frame were most relevant for deciding the next move.
2. SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations): These are commonly used model-agnostic techniques that provide feature attributions by perturbing input features and measuring their impact on the policy's decisions.
3. Counterfactual Explanations: This technique analyzes how the agent's decision would change if certain input features were modified. It helps answer questions like: *What would the agent have done if a different feature value were observed?*
4. Attention Mechanisms: Some RL models incorporate attention layers to focus on the most critical input features, thereby improving both performance and interpretability.

3. Policy-Level Explanation

While feature importance focuses on individual state-action pairs, Policy-Level Explainability provides a higher-level view of how the agent's decision-making evolves over time.

Key Aspects of Policy-Level Explanation

- Policy Visualization: Techniques such as decision heatmaps and trajectory plots allow researchers to see how the agent's policy changes in different states.
- Action Sequence Analysis: Instead of just looking at a single state-action pair, this approach evaluates how decisions unfold across multiple time steps.
- Policy Comparison (π\pi vs π′\pi'): The diagram illustrates how the agent updates its policy over time. Understanding these updates helps in analyzing:
    - Why the agent initially made certain mistakes.
    - How the learning process corrects those mistakes.
    - Whether the updated policy π′\pi' leads to more optimal behavior.
- Causal Reasoning: Some studies incorporate causal models to explain why the agent took a particular path rather than an alternative.

4. Learning Process and Markov Decision Process (MDP) (Blue Box)

The learning process is at the heart of RL, where the agent refines its decision-making based on accumulated experiences. The Experience Tuple (s, a, r, s') plays a crucial role in updating the agent's policy.

The Role of Experience Tuples

An experience tuple consists of: Current state (s), Action taken (a), Reward received (r), Next state (s′)

These tuples are stored and used to update the **policy** π′ through various RL algorithms.

Learning Process

The learning process is responsible for adjusting the policy based on collected experiences. Some common learning algorithms include:

- Q-learning: Updates the value function based on the maximum expected future rewards.
- Policy Gradient Methods (e.g., PPO, A3C): Directly optimize the policy rather than relying on value estimation.
- Actor-Critic Methods: Combine value-based and policy-based approaches for better learning efficiency.

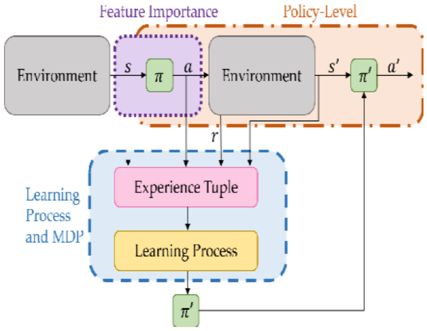

Fig.2: Explainable Reinforcement Learning Process

RESULT

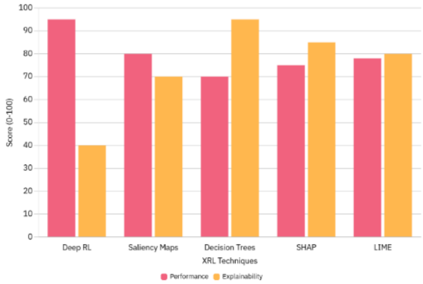

Fig.3 Performance vs. Explainability Trade-off in XRL

In Explainable Reinforcement Learning (XRL), there is an inherent trade-off between performance and explainability. Deep RL models are highly effective in complex decision-making tasks, achieving high performance, but their decision processes are often opaque, making them difficult to interpret. On the other hand, Decision Trees provide high explainability by offering clear, rule-based decision paths; however, they struggle with scalability and tend to perform poorly in high-dimensional environments. Techniques like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) serve as a middle ground, offering insights into model decisions while maintaining a reasonable level of performance. These methods help analyze feature importance and provide local explanations without compromising the overall effectiveness of RL models, making them valuable tools for balancing transparency and efficiency in real-world applications.

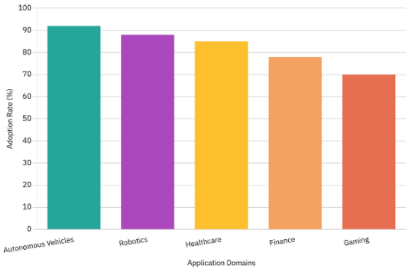

Fig.4 Adoption of XRL Techniques in Different Domains

The adoption of Explainable Reinforcement Learning (XRL) techniques varies across different domains based on their need for transparency and trust. Autonomous Vehicles (92%) and Robotics (88%) have the highest adoption rates, as interpretability is crucial for ensuring safety and reliability in automated decision-making systems. Healthcare (85%) and Finance (78%) also demonstrate strong adoption, driven by the necessity for trust, fairness, and regulatory compliance in AI-driven decision-making. In these fields, explainability helps build confidence among stakeholders by providing insights into model predictions. Gaming (70%) has the lowest adoption, primarily because performance optimization is often prioritized over interpretability. However, explainability in gaming still offers benefits, such as improving AI-driven strategies and enhancing user experience. Overall, the increasing adoption of XRL across industries highlights its growing importance in making AI systems more transparent, trustworthy, and accountable.

Table 2: Key Explainable RL (XRL) Techniques with Datasets Used

Technique	**Methodology**	**Dataset Used**	**Scope of Application**
Saliency Maps	Uses visualization techniques to highlight important input features that influence RL decisions	Atari 2600 Games (OpenAI Gym)	Gaming, Autonomous Vehicles
SHAP (Shapley Additive Explanations)	Assigns importance scores to input features to explain RL model predictions	CartPole, Healthcare RL datasets	Healthcare, Robotics, Industry 4.0
LIME (Local Interpretable Model-Agnostic Explanations)	Perturbs input features to evaluate their effect on RL agent decisions	Finance RL datasets (e.g., stock trading simulations)	Financial Markets, Algorithmic Trading
Decision Trees for RL	Converts RL policies into **human-readable decision trees**	Taxi-V3, OpenAI Gym	Finance, Healthcare, Legal AI
Attention-Based RL	Uses attention layers in deep RL networks to focus on **important state-action pairs**	MuJoCo (Robotics), Atari	Robotics, Healthcare, Autonomous Vehicles
Causal RL Explanations	Uses causal models to explain why an RL agent chose a particular action	StarCraft II, Medical Treatment Data	Healthcare, Strategy Games, Robotics
Reward Decomposition	Breaks down rewards into **interpretable sub-components** to explain RL decisions	OpenAI Gym, MuJoCo	Robotics, Industrial Automation

Conclusion

The advancements in Explainable Reinforcement Learning (XRL) have significantly improved the interpretability of RL models, addressing the long-standing issue of their black-box nature. Recent developments in feature attribution, policy visualization, causal reasoning, and reward decomposition have provided effective methods to enhance transparency while maintaining high performance. However, a key challenge in XRL remains the trade-off between explainability and model complexity. Techniques such as decision trees and rule-based RL offer clear decision paths but struggle with scalability, whereas deep learning-based RL models achieve superior performance at the cost of interpretability. To bridge this gap, hybrid approaches like attention-based models, SHAP, LIME, and saliency maps have been developed to balance transparency and accuracy.

The increasing adoption of XRL in healthcare, finance, robotics, and autonomous systems highlights its importance in domains where trust, safety, and regulatory compliance are critical. By generating human-readable explanations, XRL is enabling AI-driven decision-making in high-stakes environments. Despite these advancements, challenges remain, including the lack of standardization in evaluating explanations, the high computational cost of many XRL techniques, and the limited generalizability of explanations across different tasks. Future research must focus on developing standardized evaluation frameworks, enhancing the scalability of XRL methods, integrating human-in-the-loop learning, and leveraging causal inference for deeper interpretability. As XRL continues to evolve, it has the potential to revolutionize AI decision-making by fostering more transparent, trustworthy, and human-aligned reinforcement learning systems across multiple industries.

References

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine, 40(2)*, 44-58.

Muñoz, J., et al. (2023). Challenges in Explainable Reinforcement Learning: A Survey of Methods and Metrics. *Journal of AI Research*.

Puiutta, E., & Veith, E. M. (2020). Explainable Reinforcement Learning: A Survey. *arXiv preprint arXiv:2005.06247*.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. *MIT press*.

Zahavy, T., Ben-Zrihem, N., & Mannor, S. (2016). Graying the black box: Understanding DQNs. *International Conference on Machine Learning (ICML)*, 1899-1908.

Amir, D., Kamar, E., & Shah, J. (2019). Highlights: Summarizing agent behavior to people. *International Conference on Human-Robot Interaction (HRI)*, 458-466.

Andreas, J., Klein, D., & Levine, S. (2017). Modular multitask reinforcement learning with policy sketches. *International Conference on Machine Learning (ICML)*, 166-175.

Atrey, A., Clary, K., & Jensen, D. (2020). Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning. *NeurIPS Workshop on Interpretability*.

Christiano, P. F., Leike, J., Brown, T., et al. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Ehsan, U., Harrison, B., Chan, L., & Riedl, M. O. (2019). Automated rationale generation: Explainable AI for reinforcement learning agents. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2662-2669.

Greydanus, S., Koul, A., Dodge, J., & Fern, A. (2018). Visualizing and understanding Atari agents. *International Conference on Machine Learning (ICML)*, 1792-1801.

Juozapaitis, Z., Kazhamiaka, F., et al. (2019). Explainable reinforcement learning via reward decomposition. *International Conference on Learning Representations (ICLR)*.

Liu, Y., Doshi-Velez, F., & Brunskill, E. (2018). Toward interpretability in deep reinforcement learning. *arXiv preprint arXiv:1807.09794*.

Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2020). Explainable reinforcement learning through a causal lens. *AAAI Conference on Artificial Intelligence*, 13631-13639.