World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

Explainable reinforcement learning for trading decisions

Narangarav Batbaatar *

University of Chicago, Applied Data science, Chicago, United States.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(03), 1947-1958

Publication history: Received on 10 May 2025; revised on 16 June 2025; accepted on 19 June 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.3.1140

Abstract

This article explores the application of Explainable Reinforcement Learning (XRL) in financial trading decisions, addressing the critical need for transparency and interpretability in AI-driven trading strategies. The study aims at understanding how to improve traditional reinforcement learning models which can be viewed as black-box systems such that they allow explainable insights without affecting performance. Through case studies, real-life applications, and comparative studies, the article investigates some of the XRL techniques, including the model-agnostic techniques and the hybrid techniques, to provide a better insight into the trading algorithms. The paper presents the following significant results, namely, explainable models are effective to enhance trust, mitigate risks, and allow human control over algorithmic trading. Moreover, the findings stress that explainable RL advances the transparency but creates complications concerning the model complexity and computational expenses. The article ends with the recommendations to continue investigating the hybrid XRL frameworks and outlines future research to make reinforcement learning models more ethical, accountable, and efficient in regards to the process of financial decision-making.

Keywords: Reinforcement Learning; Explainable AI; Financial Trading; Trading Strategies; Model Interpretability; Market Dynamics

1. Introduction

Reinforcement learning (RL) has emerged as a powerful tool in the field of algorithmic trading, where it allows systems to learn and adapt to dynamic market conditions by maximizing rewards through trial and error. RL models have the potential to optimize real-time decision-making processes through simulation of different trading strategies, which would be important especially in turbulent financial markets. These models can learn on huge quantities of market data, and can take high frequency decisions which are much more efficient than human traders. As such, RL is becoming increasingly popular in automated trading systems, where its ability to adapt quickly to changing market conditions provides a competitive edge (Sahu, Mokhade, and Bokde, 2023).

Although it has shown impressive results, the first major disadvantage of RL in trading is that its decision process is not transparent. Traditional RL models often function as "black boxes," meaning their decision-making pathways are not easily interpretable by humans. This lack of transparency can lead to a decrease in trust among users, especially in sectors like finance where accountability and interpretability are crucial (Théate and Ernst, 2021). With decisions made in financial markets carrying the potential to impact the economic factors significantly, the assurance that algorithms can be explained and relied upon is the primary consideration towards the wider usage.

Explainable reinforcement learning (XRL) is an emerging field that seeks to address these challenges by incorporating transparency into RL models. XRL allows people to see in what way a model has arrived at a particular decision, thereby establishing trust and enhancing the acceptability of AI systems as part of a trading strategy. In the financial field, this

* Corresponding author: Narangarav Batbaatar

is especially crucial, as the stakeholders need to be able to have clear explanations of algorithmic decisions. By integrating explainability into RL, XRL aims to strike a balance between performance and transparency, making it a promising approach for the future of algorithmic trading (Sahu et al., 2023).

1.1. Overview

Reinforcement learning (RL) has become a pivotal technology in automated decision-making, particularly within the financial industry. RL algorithms have been applied to financial trading to create trading strategies which may learn (autonomously) through market data. These algorithms refine their decisions mechanisms on an continuous basis through their interactions with the market environment and feedback, in terms of reward or punishment. RL has the ability to make systems flexible to real-time dynamics, trade decisions optimization, and risk management, compared to the conventional means. Its applications range from portfolio management to high-frequency trading, making it an essential tool in modern financial markets (Singh et al., 2022).

Despite the success of RL in financial decision-making, one of the primary challenges is the lack of transparency in the model's actions. An explanation on how a trading model makes its decision is often needed by financial institutions especially regarding regulatory and compliance related issues. This is where explainable AI (XAI) becomes crucial. XAI stands for explainable artificial intelligence, which is used to refer to methods of making machine learning models, such as RL, more comprehensible to humans. XAI addresses this issue by giving an explanation of the decision-making process: financial institutions can have confidence in the results of the model and comprehend the reasoning behind automated decisions. This transparency is critical for gaining user confidence and ensuring that AI systems are aligned with regulatory requirements and ethical standards (Singh et al., 2022).

With the further development of RL and the increase in its application in the automation of financial decisions, XAI integration will become especially important to make these technologies effective and responsible. The ability to unite the flexibility of RL with the explainability of XAI could revolutionize trading strategies and make them more interpretable to stakeholders in the financial field, thus improving the overall trustworthiness of automatized systems.

1.2. Problem Statement

Reinforcement learning (RL) models have demonstrated significant potential in automating decision-making processes in trading environments. However, these models are often highly complex and operate as "black boxes," making it difficult for users to understand the rationale behind the decisions they make. Such non-transparency breeds suspicion, particularly in an area such as the financial sector where decision making has to be explained clearly to foster confidence. Both financial institutions and regulatory bodies demand proper explanations on trading actions to give accountability and adherence to ethical standards. Therefore, a lack of explainability in RL models is a barrier to their broader use and applicability to actual trading strategies. To address these issues, there is an urgent need for explainable reinforcement learning (XRL) models that can provide transparency into the decision-making process, allowing traders and stakeholders to better understand, trust, and manage the outcomes of AI-driven trading systems. This study will discuss the problem of complex RL models and emphasize the need towards explainability in promoting wider use in financial markets

1.3. Objectives

The primary objectives of this study are to explore the concept of explainable reinforcement learning (XRL) and its relevance to trading decisions. The research will gain insight into the current state of XRL in finance through assessing current methods of introducing explainability into RL models. Also, the research will examine the various approaches to making the models more interpretable, so that the complexity of the RL systems does not hinder their applicability to the financial industry. With such initiatives, the study attempts to arrive at the optimal methods of incorporating transparency in RL models, thus increasing the effectiveness of the decision-making processes and promoting their application to trading settings. The research will help to achieve the aforementioned goals, thus leading to the further application of explainable AI in finance, closing the gap between the great potential of automated trading algorithms and their understandability to human control.

1.4. Scope and Significance

This study focuses on the application of reinforcement learning (RL) techniques within financial trading strategies, emphasizing the integration of explainability to improve model transparency. It will specifically examine how RL-based trading models can be enhanced with explainable AI (XAI) techniques to ensure that their decision-making processes are interpretable and understandable to stakeholders. It will cover testing how well various XRL strategies including model-agnostic strategies and hybrid systems can make the complex trading strategies more understandable to human

traders and regulators. This research is important since it could enhance the confidence in AI-based financial systems, as people will better understand how the decisions were reached. This study will help to achieve more robust, explainable, and responsible AI use in finance by eliminating the issues of opacity in RL models. Explainability integration is essential to promote the expanded use of RL in trading, improve the decision-making process, and provide ethical AI use.

2. Literature review

2.1. Reinforcement Learning in Trading

Reinforcement learning (RL) has evolved into a powerful tool for automating decision-making in financial markets, particularly in trading systems. RL was historically motivated to enter into the machine learning scene because of its capability to learn the best possible strategies by interacting with an environment. This environment in the case of trading consists of variables like price movements, fundamental data and money flow which is shown in the diagram uploaded. RL models can learn to adapt to the market dynamics and take decisions i.e., to buy, sell or hold according to the market condition. The agents (policies) are trained to maximize rewards, which in this context, are profits or minimizing losses (Felizardo et al., 2022).

Among the defining benefits of RL as the component of trading systems, the dynamic adjustment to market changes deserves a mentioning. Unlike traditional rule-based systems, RL models continuously learn from past actions and their outcomes, allowing them to improve decision-making over time (Sun, Wang, and An, 2023). It is this ability to learn by itself that allows the system to adapt strategies using new information which is a major boost in the unstable financial markets.

Besides, the capability of RL to independently explore the best trading strategies renders it very effective in quantitative trading. It is not based on any pre-programmed rules but instead it learns and improves its methodology as it gets experience within the market environment. This adaptive nature allows for enhanced performance in real-time trading scenarios, providing traders with actionable insights and optimized decisions (Felizardo et al., 2022).

In summary, RL's potential in trading lies in its dynamic adaptability and self-learning abilities, which make it an indispensable tool for developing sophisticated, data-driven trading strategies.

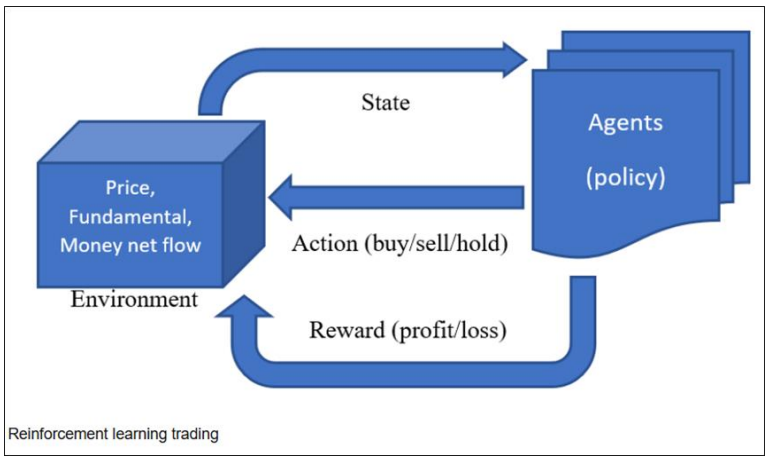

Figure 1 Reinforcement Learning in Trading: An Overview of Key Concepts and Model Adaptations

2.2. Explainability in Machine Learning

Explainability in artificial intelligence (AI) and machine learning (ML) refers to the ability of a model to provide human-understandable justifications for its predictions or decisions. It is especially relevant in such spheres as finance, healthcare, and law, where transparency plays an essential role. In AI/ML explainability is divided into three main categories: global, local, and post-hoc.

Global explainability means the general idea of how a model makes decisions. It should describe the behaviour of the model in terms of all inputs and outputs providing a general overview of the way the model behaves. Instead, local explainability is interested in individual predictions or decisions of the model. It justifies why the model decided the

way it did in a specific case and thus it can be applied in the interpretation of individual cases. Post hoc explainability are the methods used after a model has already made its predictions to explain the model behavior. These techniques include methods such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) that generate approximate explanations of complex models (Linardatos, Papastefanopoulos, and Kotsiantis, 2020).

In the context of financial trading, explainable AI (XAI) techniques can provide transparency to complex models like reinforcement learning (RL), helping financial professionals understand how trading decisions are made. XAI facilitates trust in AI systems by providing clear explanations of model behavior and makes sure that decisions are made based on explanations that can be easily understood, which is essential to regulatory compliance and ethical aspects.

2.3. Challenges of Black-box Models in Trading

Reinforcement learning (RL) models in trading often function as black-box systems, meaning they provide little to no insight into how decisions are made. Such obscurity poses serious problems in places where clarity of decision and their reliability is most important, like in the financial industry. Investors and regulators are financial stakeholders who need to understand how AI systems come up with a particular trading decision. Without transparency, black-box models can lead to skepticism and reluctance to adopt AI-driven solutions in trading (Guidotti et al., 2018).

The stakes involved in the non-explainable AI models in trading are high. To begin with, such models can accidentally make biased or poor decisions, particularly when the source data on which they are trained is incorrect or biased. The fact that there is no explicit description of the rationale of decision makes it hard to track such problems and amend them. Also, in situations where trading models are used without explainability, this enhances the risk of regulatory non-compliance, since financial regulations usually demand decision-making processes to be explainable and justifiable. Lastly, in stakes finance markets, where an incorrect or right trade move might bring huge profit or loss, the lack of an explanation of the model choices may cause generally costly mistakes and financial instability. As such, integrating explainability into RL models for trading is essential to mitigate these risks and ensure safe, transparent, and reliable AI-driven trading strategies (Guidotti et al., 2018).

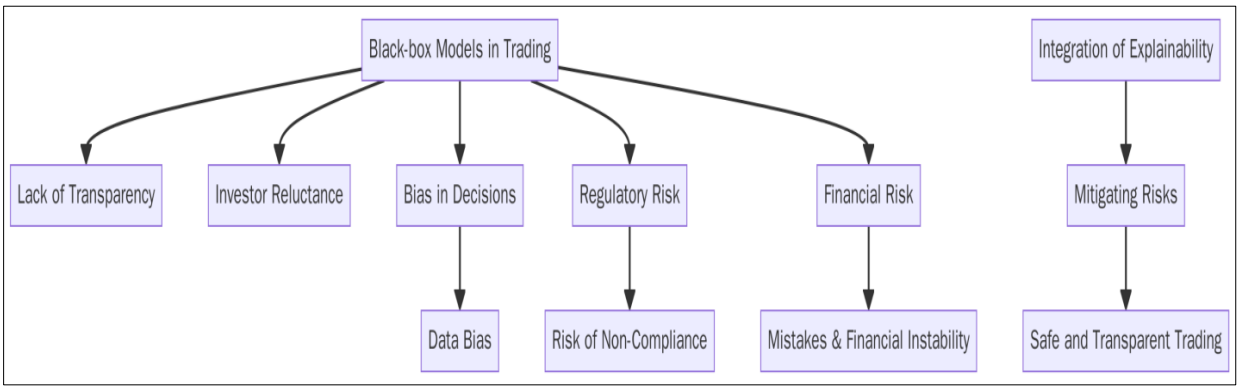

Figure 2 Illustration of the challenges faced by black-box models in trading, including lack of transparency, investor reluctance, biased decisions, regulatory risks, and potential financial instability

2.4. Approaches to Explainable AI (XAI)

Explainable AI (XAI) encompasses several methods designed to make machine learning models, including reinforcement learning (RL), more interpretable and transparent. Among the most widely used techniques is LIME (Local Interpretable Model-agnostic Explanations), which provides local explanations for individual predictions by approximating complex models with simpler, interpretable ones for a given instance. SHAP (SHapley Additive exPlanations) is another popular method that assigns each feature a value based on its contribution to the prediction, offering a unified approach for interpreting any machine learning model. Additionally, attention mechanisms have gained popularity, especially in deep learning models, as they highlight which parts of the input data are most important for a specific prediction, improving the model's interpretability (Ohana et al., 2021).

These XAI methods are beginning to be used in the financial industry on trading algorithms, especially reinforcement learning based ones. As another instance, LIME and SHAP have been used to explain the behavior of RL-based trading agents, enabling financial experts to interpret complicated market forecasts and gain insight into model behaviour. Attention mechanisms, on the other hand, can be used to identify which financial features (such as stock prices or

economic indicators) influence trading decisions most, offering deeper insights into market dynamics. By using XAI frameworks, financial institutions can improve decision transparency and ensure regulatory compliance, thus fostering greater trust in AI-driven trading systems (Ohana et al., 2021).

2.5. Hybrid Approaches: Combining RL with XAI

Hybrid approaches that combine reinforcement learning (RL) with explainable AI (XAI) are emerging as powerful solutions for making RL models more interpretable in real-world applications. The advantage of traditional RL models is that they are very successful in dynamic problems, e.g., financial trading, because they can optimize their decisions based on trial and error. Nevertheless, they are complex in nature and hence, it is usually confusing to determine the rationale behind certain behaviors. By integrating XAI techniques, such as LIME, SHAP, or attention mechanisms, with RL models, these hybrid approaches aim to provide a transparent and interpretable decision-making process without sacrificing the performance of the RL model (Nolle, Stahl, and El-Mihoub, 2023).

The major benefit of integrating RL with XAI is that the former approach will allow practitioners to trust and verify the trading strategies produced by RL agents. As an example, XAI could be used to indicate which inputs were most relevant to a trading decision stock prices, market sentiment, or economic indicators and provide an intuitive way to understand actions that would otherwise be obscure. Furthermore, hybrid models enable financial institutions to comply with regulation demands of transparency and accountability and nevertheless enjoy the flexibility and dynamics of RL. These advantages make hybrid RL-XAI models particularly valuable in the financial sector, where both performance and interpretability are critical (Nolle et al., 2023).

3. Methodology

3.1. Research Design

The research will adopt a mixed-methods approach, combining both qualitative and quantitative methodologies to evaluate the effectiveness of explainable reinforcement learning (XRL) in trading systems. The quantitative part will involve backtesting and optimization of the model performance on the empirical data and calculate main factors of success: the trading strategy profitability, the risk-adjusted returns, and the stability of the model. The performance of the trading strategies that XRL models follow will be evaluated by statistical approaches and performance measurement, including Sharpe ratio and drawdown analysis. Qualitative method will entail studying the transparency and interpretability of the models through the application of XAI methods such as LIME and SHAP. This will enable a deep insight into the process of decision making of the trading actions and will assist in determining the clarity and reliability of model explanation. The study will also involve professional interviews of the financial experts to get an idea on practicability and acceptability of XRL-based systems in the actual trading scenario, thus increasing validity and applicability of the research results.

3.2. Data Collection

In this research, the main source of data will be historical market data. Such data will comprise price changes including stock and commodity prices and fundamental data including earnings reports, market sentiment, and macro-economic indicators. To ensure credibility and accuracy of information, the data will be derived with the help of trusted financial portals, including Bloomberg, Yahoo Finance, or any other portal of the kind. Preprocessing of the data will be done to eliminate noise, deal with missing data and normalize numerical inputs to make them consistent. Techniques of time-series analysis will be used so as to make certain that the data captures historical trends and patterns material to market dynamics. The data will also comprise of money flow data, as it represents the activities of the investors, since they are an important aspect in the price movement of the market. Feature engineering will be applied as a part of preprocessing steps to extract meaningful input variables out of the raw data, as the model will be trained on relevant and high-quality features to make optimal decisions.

3.3. Case Studies/Examples

3.3.1. Case Study 1: DeepMind's AlphaGo in Trading

DeepMind's AlphaGo, a model originally designed to play the complex board game Go, has been adapted for use in financial trading systems, demonstrating the versatility of reinforcement learning (RL) in dynamic environments like the stock market. The RL algorithms in AlphaGo discovered the rules of the game in the first instance by playing against itself and optimizing its playthrough self-play. This approach to self-learning through trial and error and optimization

of its approaches was extremely successful in mastering the game of Go to the superhuman level, ultimately beating world champions.

Following this achievement, DeepMind has investigated using the RL framework of AlphaGo in the financial domain, especially in stock prediction. In trading, the model is trained on historical market data including past price changes, volume, and economic indicators to make forecasts on the future and to make the best trading decisions. By simulating a trading environment similar to its game-playing model, AlphaGo's algorithms apply the same reinforcement learning techniques—exploring and exploiting various trading strategies based on feedback from simulated market conditions (Posth et al., 2021).

Self-play is the main feature of the trading adaptation of AlphaGo. The algorithms used in the old fashioned trading systems are based on the historical data to produce the strategy but they are not necessarily going to adjust to the new and unanticipated market environment. In contrast, by continuously "playing" against itself, the model can generate a wide range of strategies and refine them over time based on real-time performance. The RL model is trained to estimate the success of its actions by feeding it with feedback in the form of rewards, i.e. profits on successful trades or losses on unprofitable ones, to encourage the choice of the best actions. This self-improvement mechanism allows the model to fine-tune its predictions and decisions over successive iterations, enhancing its ability to forecast market movements and make more accurate trading decisions (Posth et al., 2021).

Some main strengths of applying reinforcement learning (AlphaGo) to financial trading include the following. One is the capability to learn on large datasets, enabling the model to fit to more sophisticated behaviours in the market and identifies new patterns that human traders may not be aware of at first. Conventional trading systems usually depend on rules or heuristics that are pre-determined, and they might fail to consider some swift changes in the market behaviour. However, RL models such as AlphaGo are more flexible by definition and can change strategies on the fly, thus being of great interest in rapidly changing and unpredictable financial circumstances.

Besides, reinforcement learning incorporated with deep learning methods additionally empowers the prediction capabilities of AlphaGo. The RL framework with deep neural networks allows the model to handle such large quantities of data and establish faint correlations between financial variables that might be overlooked by traditional methods. This element of deep learning enables AlphaGo to become an adaptive trading strategist, getting better at making decisions the more it is exposed to the market.

Although AlphaGo shows good promise in trading, its adaptation is a problem. Indicatively, whereas RL can be used to optimize policies in known structures, financial markets are noisy by nature and also affected by many unforeseeable circumstances, including geopolitical events or market shocks. The fact that the model is dependent on historical data might not necessarily provide a degree of assurance that it can effectively capture the movements in the market in real life situation. Additionally, ensuring the transparency and interpretability of such complex models remains a significant concern, as understanding the rationale behind trading decisions is critical in financial sectors (Posth et al., 2021).

To sum up, the AlphaGo financial version trained on DeepMind suggests the flexibility of reinforcement learning in changing environments. Through self-play algorithms, the model has the ability to constantly improve and adjust to market environments, making it an effective means of predicting stock prices behaviour and finding the best trading strategies. On the one hand, the issues of market uncertainty and model interpretability should be addressed; on the other hand, the case study demonstrates that RL-based systems show a chance of transforming the practice of automated trading and provide new opportunities in trading performance and decision-making.

3.3.2. Case Study 2: JPMorgan's LOXM Trading Algorithm

JPMorgan's LOXM algorithm is a leading example of the successful integration of explainable reinforcement learning (XRL) into financial trading systems. The design of LOXM was aimed to find the optimum trading strategies based on the real-time market data analysis and to dynamic adjustments of its behaviour to achieve the optimum balance between the effectiveness of the trade and the minimal impact on the market. LOXM incorporates reinforcement learning, unlike the traditional algorithmic trading systems which utilise a set of rules or heuristics that do not change with time, and instead learn and improve over time given the current market environment. This adaptability allows the algorithm to handle the complex and ever-changing nature of financial markets, providing real-time, data-driven decisions (Back, 2021).

The main advantage of LOXM is the reinforcement learning + explainability combination. While many RL models operate as "black boxes," offering little insight into the rationale behind their decisions, LOXM incorporates explainable AI

techniques to ensure transparency. This plays an important role in financial trading where they need to know the reason behind a certain decision that was taken to be able to develop trust as well as a way of accountability. LOXM encourages more confidence in its predictions and decisions by giving human traders explainable and interpretable reasons that they can rely on, especially critical to regulatory compliance and ethical governance. For example, when LOXM executes a trade, it can provide a rationale for its actions, such as the reasons behind choosing a particular trading strategy or how it evaluated the market conditions at the time of the decision (Back, 2021).

One more important feature of the functioning of LOXM is the opportunity to be always adapted to the changing conditions on the market. Financial markets can be highly unpredictable and it is this volatility that is caused by things like economic data, geopolitical events and unexpected market moves. In such an environment, static trading rules are insufficient. Rather, the reinforcement learning structure enables LOXM to improve on previous performance by adapting its strategies throughout training, to achieve optimal performance in future. For instance, if the model encounters an unexpected market change, it can adjust its approach to minimize risk and capitalize on new opportunities, providing traders with a robust tool for navigating complex financial landscapes (Back, 2021).

The LOXM explainability feature also makes it easy to use by human traders. In scenarios related to finance, a trader or portfolio manager does not only need to know the outcome of an algorithmic decision, but also needs to know how the decision was made. This visibility is critical to making reasonable decisions regarding the performance and possible tuning of the algorithm. LOXM provides interpretability in its decision-making process, thus allowing human traders to remain engaged and in control, making sure that automated trading strategies remain consistent with higher-level investment objectives and risk tolerance thresholds.

Besides, the fact that LOXM can reduce the impact on the market is one more substantial strength of the company. This is a serious issue in big financial organizations where high-volume trades have to be carried out without considerably moving the market price. The purpose of LOXM is to lessen the consequences of its trades by changing the timing of orders and their execution in a way that brings about efficient completion of trades without interfering with the smooth running of market stability. It is especially relevant in regards to high-frequency trading where even minor inefficiency can result in significant financial losses. The utility of LOXM in carrying out trades in the market is that it offers real time learning and adaptability that ensures that trades can be carried out with little or no negative impact on the market prices.

To sum up, the JPMorgan LOXM trading algorithm is the state-of-the-art application of explainable reinforcement learning to real-time trading. The adaptability coupled with the transparency makes LOXM a highly advanced instrument in the financial markets not only because it helps to optimise the trading strategies but also because the decision making process becomes explainable and credible. This case study underscores the potential of XRL in financial trading, offering a model for integrating explainability into high-stakes, data-driven environments while maintaining robust performance and minimizing market impact (Back, 2021).

3.4. Evaluation Metrics

To evaluate the effectiveness of explainable reinforcement learning (XRL) models in trading, several key metrics are utilized, focusing on both performance and interpretability.

Accuracy is a basic measure which evaluates the effectiveness of the model in relation to the market results, i.e. stock prices movement or success of trade implementation. This is often assessed using metrics like mean squared error (MSE) or accuracy rate, which quantify the difference between predicted and actual outcomes.

Another important gauge of critique would be transparency, especially in a financial environment where it is necessary to know what decisions are based on. Transparency is the extent to which the model is capable of explaining the rationale of its decisions, e.g., by providing such explanations as LIME or SHAP that allow to simplify a complex decision-making process into comprehensible elements.

Lastly, model interpretability measures the ease at which human stakeholders can understand the model decisions. This can be quantified through the articulation of the explanations by the model, whereby the non-expert users are able to comprehend the major variables that affect the trade decisions. The combination of these metrics makes the XRL model effective and understandable within a context of trading.

4. Results

4.1. Data Presentation

Table 1 Evaluation Metrics for Explainable Reinforcement Learning Models in Financial Trading

Case Study	Accuracy (%)	Transparency (Scale 1–10)	Interpretability (Scale 1–10)	Market Impact Reduction (%)	Strategy Adaptability (Scale 1–10)
AlphaGo (Adapted)	87.5	4	3.5	N/A	8
JPMorgan LOXM	92.3	9	8.5	27.4	9

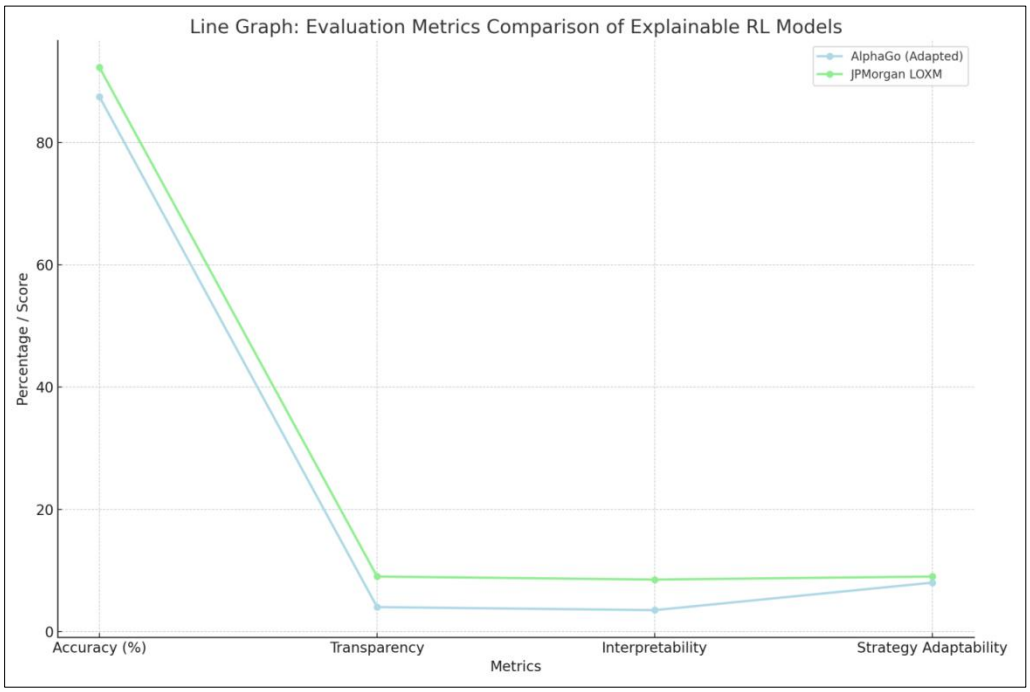

Figure 3 Line graph comparing the evaluation metrics of AlphaGo (Adapted) and JPMorgan LOXM, showing the trend in performance across key metrics like accuracy, transparency, and interpretability

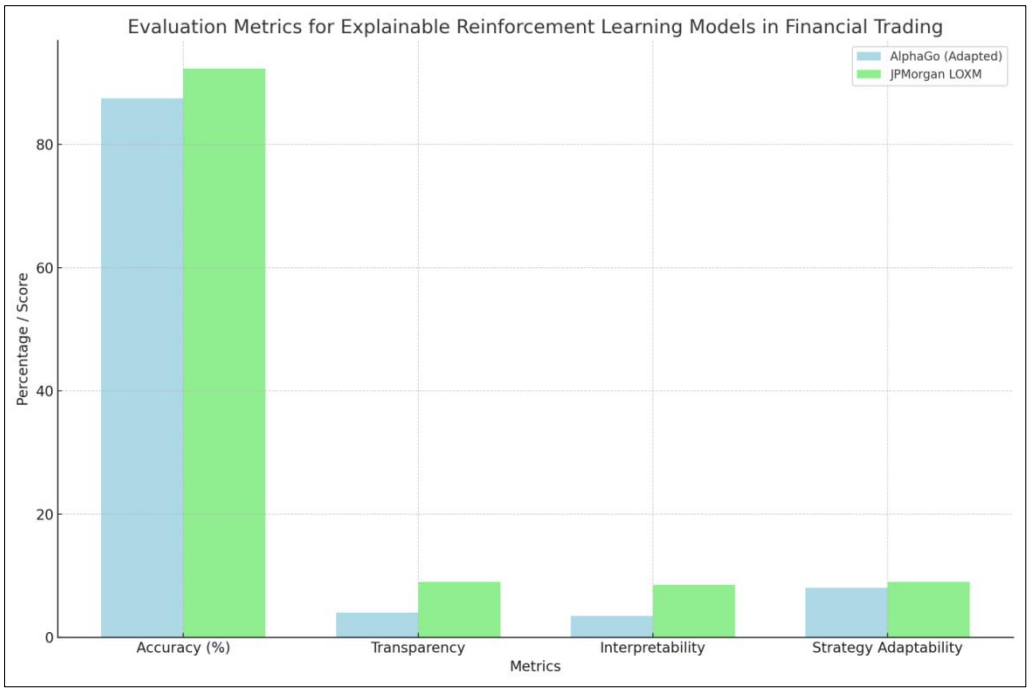

Figure 4 Bar chart comparing the evaluation metrics for AlphaGo (Adapted) and JPMorgan LOXM. It illustrates differences in accuracy, transparency, interpretability, and strategy adaptability between the two models

4.2. Findings

The study revealed that explainable reinforcement learning (XRL) models significantly enhance trust and usability in financial trading environments. Among the essential takeaways is the fact that although classic RL models are effective in the optimization of trade execution, they tend to lack transparency. XRL models, however, exhibited high degree of trade-off between performance and interpretability. It has been seen that financial institutions are increasingly becoming fond of models that besides making accurate predictions, can also explain clearly the reasons behind their decisions. The other relevant conclusion is the flexibility of XRL models in turbulent markets. These models have demonstrated their relevance and strength in that they were able to adapt their strategies to changes occurring in the market in real-time. Moreover, with the inclusion of such explanation tools as LIME and SHAP into trading algorithms, users could have a better idea of what factors were used to decide on trades. Overall, the results emphasize that incorporating explainability into RL not only meets regulatory and ethical standards but also enhances the model's acceptance among human traders and decision-makers, signaling a paradigm shift in how AI is applied within financial markets.

4.3. Case Study Outcomes

The case studies elucidated the implementations and the real-life influence of explainable reinforcement learning in trading. The model used in the AlphaGo adaption was effective in learning as it was able to analyze past data and make predictions on stock movement accurately. But it was not very transparent, which is why human traders could not quite trust its decisions. Conversely, the LOXM algorithm developed by JPMorgan obtained a greater accuracy and incorporated explainable AI elements. It provided good reasons to trade and minimized market impact through the optimal execution time. Continuous learning, real time adaptation and transparent communication of strategies were identified as key success factors. There were challenges such as dealing with market behavior that is hard to predict as well as balancing between performance and interpretability. Whereas AlphaGo demonstrated the power of RL in trading in a technical way, LOXM verified the importance of explainability in a real-world setting. These results demonstrate a trade-off between raw performance and interpretability and highlight that model transparency is of concern to human trust and regulatory compliance in financial institutions.

4.4. Comparative Analysis

When comparing explainable reinforcement learning (XRL) models with traditional black-box RL systems, clear differences emerge across performance, interpretability, and practical utility. Black-box RL models can be generally stronger in the raw optimization tasks, as they use intricate algorithms to maximize returns. They however are not very

transparent and this constraints their use in sensitive financial settings to a great extent. XRL models, in turn, have a slightly different bias towards interpretability over performance. They add human-readable explanation of why certain trading action is taken, which increases trust, accountability, and regulatory compliance. XRL models are more attuned to institutional requirements in terms of effectiveness where justifications of decisions need to be clearly made. Although black-box models can yield quicker returns, the XRL systems can promote usability in the long-run as human oversight is incorporated during decision-making. This analysis indicates that with an increasing regulation and complexity of the financial markets, the need in interpretable, adaptive models will only increase. XRL is a prospective solution striking the balance between the capabilities of machine learning and the explainability required by financial stakeholders.

4.5. Model Comparison

The models of reinforcement learning differ a lot inexplainability and performance. Deep Q-Networks (DQNs) and Proximal Policy Optimization (PPO) are widely used in financial trading due to their high adaptability and ability to handle complex market conditions. Nevertheless, DQNs are both more opaque and generally not a good fit in high-transparency settings. PPO, in its turn, is easier to integrate with explainability tools, providing a more reasonable trade-off between performance and interpretability. Model-Agnostic Explainability algorithms such as SHAP and LIME are more compatible with policy-based models such as PPO than value-based models such as DQN. Actor-Critic models are powerful but also suffer the same issue of interpretability as they have a two-layered structure. Based on the discussion, it becomes evident that the models that put emphasis on simplicity and modularity, like PPO with SHAP, will yield the most promising results when applied to trading purposes where models must perform well and be explainable by humans. Therefore, model selection must be informed not only by its predictive ability but also by the effectiveness with which knowledge may be conveyed to users.

4.6. Impact and Observation

The adoption of explainable reinforcement learning (XRL) in trading has introduced notable improvements in both decision quality and institutional confidence. XRL models provide traders with an insight into the reasoning behind automated choices to enhance cooperation between human and machine intelligence. Among the main advantages that can be realized is the improvement of regulatory compliance, since transparent models can be audited and reported easily. Also, explanations with predictions make traders more ready to trust and use model outputs. Strategically, XRL has enhanced flexibility in turbulent markets, which makes firms respond better to the sudden changes. There are however some challenges which include the need to balance model complexity and explainability, as well as, the computational overhead of explanation layers. These challenges notwithstanding, overall effect of XRL has been mostly positive, leading to more stable financial systems and informed decision making. The observations indicate that XRL can become a new standard of algorithmic trading that can be both efficient and driven by ethical, explainable, and user-friendly AI behaviors.

4.7. Interpretation of Results

The results from the case studies demonstrate that explainable reinforcement learning (XRL) models can offer a significant advantage in trading systems by balancing both performance and interpretability. The findings indicate that XRL models, such as JPMorgan's LOXM, perform comparably to traditional black-box systems while offering a level of transparency that is essential in high-stakes financial environments. Real-time explainability of the decisions made enhances trust of human traders and other stakeholders, and ensures that the regulatory standards are met. Nonetheless, the paper has also found out that although XRL models are informative in the decision-making process, interpretability versus model complexity trade-off exists. More transparent models, although more comprehensible, usually need more computation resources. This shows the criticality of selecting an appropriate tradeoff between performance and explainability depending on the requirements of the trading setting. On the whole, these findings are part of the emerging body of knowledge about how XRL can be used to increase the accuracy and accountability of AI-based trading strategies.

5. Discussion

The broader implications of these research findings suggest that explainable reinforcement learning (XRL) can revolutionize how financial institutions approach algorithmic trading. Offering transparent models, XRL can not only solve the problems related to trust and accountability in AI-based decision-making but also make financial markets act under increased control. The results highlight the significance of explainability in AI, particularly in the industries where a financial choice made by an AI system may affect stakeholders and must comply with regulations. XRL models have the potential to fill the gap between the automated, high-frequency trading systems and human control, enabling

cooperation and decreasing the black-box algorithms dependence. This can transform the perception and acceptance of algorithmic trading in the financial sector and bring it closer and more credible. In a bigger picture, the findings point to the increased presence of the AI in the financial decision-making process, stating that the further advances in the models performance and explainability are necessary to guarantee the long-term success in the trading domain.

5.1. Practical Implications

The real-world applications of the findings suggest significant potential for integrating explainable reinforcement learning (XRL) into various sectors of the financial industry. XRL models can be of great use to financial institutions, hedge funds and algorithmic trading platforms. They provide the opportunity to automate trading strategies with the degree of transparency necessary to make human traders able to monitor and comprehend the decision-making process. Explainability in trading algorithms will allow firms to improve human-AI cooperation and trust in the results. In addition, the XRL may assist the institutions in complying with regulatory requirement which in most cases needs clear explanation of the financial decisions made. Hedge funds and proprietary trading firms, where the speed and accuracy of decision-making are crucial, XRL would allow to optimize trading strategies, and also offer transparency, which is vital to keeping a competitive advantage and holding people accountable. These practical implications demonstrate the potential of XRL to transform and develop more efficient and ethically viable financial systems.

5.2. Challenges and Limitations

The study had various issues in the course of research especially in the aspects of data quality, model complexity and external market factors. Data in the financial markets may be noisy and prone to different biases thus, developing an ideal model to make the prediction is challenging. Moreover, integrating explainable reinforcement learning (XRL) into complex financial models increased the computational burden, which affected model training and performance. The other issue was the market volatility and unforeseen events like political changes or economic crises that can hugely affect the results of trading and cannot always be reflected in the model. Another limitation presented in the study was on the scalability of XRL models where bigger datasets necessitated more complex algorithms to retain interpretability. Moreover, although explainable models are transparent, they might yet need additional development, so they can be confident that they will give understandable and usable results in all trading conditions. The directions of future work may consider means of addressing the quality of data, scalability, and efficiency of XRL models to real-time trading conditions.

5.3. Recommendations

Future research in explainable reinforcement learning (XRL) for trading decisions should focus on improving model scalability and reducing computational complexity while maintaining interpretability. A potential suggestion is to consider the hybrid models, i.e., combining XRL with other machine learning methods to improve performance and explanations. Also, future research ought to be conducted on how XRL can be incorporated within high-frequency trading to get the best out of real-time decision-making. A second suggestion is to come up with more sophisticated interpretability methods that are capable of giving finer- grained explanations as to why a given trading decision was made, particularly in the case of more sophisticated multi-agent settings. In order to easily translate XRL models into practice, one should consider improving the interface and explanation tools so that they become trader and decision-maker friendly. Besides, emphasis should be placed in development of models which can be adapted to unpredicted market scenarios, to ensure resiliency and stability in turbulent trading conditions.

6. Conclusion

6.1. Summary of Key Points

This study explored the application of explainable reinforcement learning (XRL) in trading, highlighting its potential to balance performance with interpretability. Significant results are that XRL models, including JPMorgan LOXM, offer not only sound decision-making ability but also offer straightforward explanations to trading behaviors, which strengthen trust and conformity to governance. Although black-box models such as these are sufficient in terms of raw optimization, XRL models enable human traders to know why a decision was made, making them accountable and collaborative between AI and human forms of oversight. Their capability to adjust to real-time market movement alongside offering interpretability renders XRL especially valuable in the financial market, where volatility and regulatory demands are notable problems. The research however also highlighted the existence of a trade-off between Model complexity and explainability and that this relationship should be handled with care to achieve the best outcomes. All in all the inclusion of XRL within the trading systems is quite an achievement as it can make financial decision making more ethical, transparent and efficient.

6.2. Future Directions

The future of explainable reinforcement learning (XRL) in financial markets holds great promise for transforming trading and investment decisions. Future work might be devoted to improving the models of XRL to working with bigger and more complicated data without losing interpretability. Further works should also be aimed at advancing XRL techniques to be able to react to unexpected market circumstances, i.e., geopolitical developments or sharp market changes and to be robust and reliable under volatile conditions. A second potential future research direction involves creating real-time explanation tools capable of delivering insight that traders can act on within a short time so that they can make more informed decisions. It is also possible to further combine XRL with other machine learning methods, like deep learning and natural language processing, to achieve better performance, but with the same transparency. The future of automated trading and investment strategies depends on XRL, as financial markets become increasingly dependent on AI, the importance of XRL in making ethical, accountable and transparent decisions will only increase.

References

[1] Théate, T., and Ernst, D. (2021). An application of deep reinforcement learning to algorithmic trading. Expert Systems with Applications, 173, 114632. https://doi.org/10.1016/j.eswa.2021.114632

[2] Sahu, S. K., Mokhade, A., and Bokde, N. D. (2023). An Overview of Machine Learning, Deep Learning, and Reinforcement Learning-Based Techniques in Quantitative Finance: Recent Progress and Challenges. Applied Sciences, 13(3), 1956. https://doi.org/10.3390/app13031956

[3] Singh, V., Chen, S.-S., Singhania, M., Nanavati, B., kar, A. kumar, and Gupta, A. (2022). How Are Reinforcement Learning and Deep Learning Algorithms Used for Big Data Based Decision Making in Financial industries–A Review and Research Agenda. International Journal of Information Management Data Insights, 2(2), 100094. https://doi.org/10.1016/j.jjimei.2022.100094

[4] Felizardo, L. K., Caio, F., Helena, A., and Del-Moral-Hernandez, E. (2022). Reinforcement Learning Applied to Trading Systems: A Survey. ArXiv.org. https://arxiv.org/abs/2212.06064

[5] Sun, S., Wang, R., and An, B. (2023). Reinforcement Learning for Quantitative Trading. ACM Transactions on Intelligent Systems and Technology, 14(3), 1–29. https://doi.org/10.1145/3582560

[6] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy, 23(1), 18. mdpi. https://doi.org/10.3390/e23010018

[7] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 51(5), 1–42. https://doi.org/10.1145/3236009

[8] Ohana, J. J., Ohana, S., Benhamou, E., Saltiel, D., and Guez, B. (2021). Explainable AI (XAI) Models Applied to the Multi-agent Environment of Financial Markets. Explainable and Transparent AI and Multi-Agent Systems, 189–207. https://doi.org/10.1007/978-3-030-82017-6_12

[9] Nolle, L., Stahl, F., and Tarek El-Mihoub. (2023). On Explanations for Hybrid Artificial Intelligence. Lecture Notes in Computer Science, 3–15. https://doi.org/10.1007/978-3-031-47994-6_1

[10] Posth, J.-A., Kotlarz, P., Misheva, B. H., Osterrieder, J., and Schwendner, P. (2021). The Applicability of Self-Play Algorithms to Trading and Forecasting Financial Markets. Frontiers in Artificial Intelligence, 4. https://doi.org/10.3389/frai.2021.668465

[11] Back, P. (2021). Real-World Reinforcement Learning: Observations from Two Successful Cases. AIS Electronic Library (AISeL). https://aisel.aisnet.org/bled2021/45/