International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 4, July-Aug-2026, ISSN (Online): 2395-566X

Explainable Reinforcement Learning for Real-Time Credit Risk Decisioning in Wholesale Banking Ecosystems

Sakthi Sankara Balaji Sathyamurthy
Truist Financial Corporation, USA

Abstract — Wholesale banking credit decisions involve large and interconnected exposures, evolving borrower conditions, collateral movements, covenant compliance, and changing market environments. Conventional credit-risk models are commonly designed for one-time prediction tasks and may not adequately support sequential decisions such as adjusting exposure limits, revising pricing, requesting additional collateral, or escalating cases for review. This study proposes an explainable reinforcement learning framework for real-time credit-risk decisioning in wholesale banking ecosystems. The framework represents credit management as a sequential decision problem in which an agent observes dynamic borrower, facility, portfolio, and market conditions before recommending appropriate risk actions. It integrates offline reinforcement learning with policy constraints to reduce the risk of unsafe or non-compliant recommendations in regulated banking environments. The proposed model incorporates an explainability layer that provides feature-based explanations, reward decomposition, policy rationale, and counterfactual action analysis. These outputs are intended to clarify why a specific credit action is recommended, which risk factors influenced the decision, how the recommendation aligns with risk appetite, and what changes in borrower conditions could lead to an alternative outcome. The study adopts a design science and empirical evaluation approach using historical wholesale credit data, with performance assessed against conventional statistical models, machine- learning models, rule-based decision engines, and non-explainable reinforcement-learning approaches. Evaluation criteria include credit-risk performance, expected-loss control, risk-adjusted return, decision latency, policy compliance, robustness, and explanation quality. The study contributes a structured framework for applying explainable reinforcement learning to high-impact credit decisions. It also provides guidance for controlled deployment through offline training, human oversight, model validation, audit logging, and continuous monitoring. The proposed approach may strengthen the timeliness, consistency, transparency, and governance of credit-risk decisioning across wholesale banking operations.

Keywords— Explainable reinforcement learning; wholesale banking; credit risk decisioning; real-time credit risk; offline reinforcement learning; explainable artificial intelligence; model risk management; credit governance; counterparty risk; risk-adjusted return.

I. INTRODUCTION

Wholesale banking credit decisions are increasingly shaped by rapidly changing borrower conditions, facility utilisation, collateral movements, covenant performance, sector exposures, and broader market conditions. Unlike retail lending, wholesale credit management involves large and concentrated exposures, complex financing structures, negotiated terms, and significant interdependence between borrowers, counterparties, and financial markets. These characteristics require banks to make timely decisions on credit approval, exposure limits, pricing, collateral requirements, monitoring intensity, and escalation pathways.

Conventional credit-risk models remain important in wholesale banking because they support borrower rating, probability-of-default estimation, expected-loss calculation, and capital allocation. However, many of these models are designed for single-period prediction rather than sequential decision-making. A credit decision made today may affect future exposure, borrower behaviour, facility performance, and portfolio risk. For example, reducing an exposure limit, requesting additional collateral, or placing a borrower under enhanced monitoring may improve future loss outcomes but may also affect client relationships and revenue generation. Static models do not always capture these long-term consequences.

Machine-learning methods have improved the ability to identify nonlinear relationships within credit-risk data. Tree-based models, ensemble approaches, and deep-learning techniques can process a wider range of financial, behavioural, and market indicators than traditional scorecards. Nevertheless, these methods commonly provide a prediction rather than a recommendation for the next credit action. Recent reviews show that machine learning has strengthened credit-risk forecasting, but concerns remain regarding transparency, fairness, model validation, data quality, and practical deployment in regulated financial institutions (Bussmann et al., 2021; Montevechi et al., 2024).

Reinforcement learning provides a different approach by treating credit management as a sequence of decisions made over time. Instead of estimating only whether a borrower may default, a reinforcement-learning agent can learn which action is likely to produce the most favourable long-term risk-adjusted outcome under changing credit conditions. Such actions may include maintaining or reducing a facility limit, adjusting credit pricing, requesting collateral, triggering enhanced monitoring, or escalating a case for manual review. Earlier work has shown that reinforcement learning can support dynamic credit-score threshold optimisation, while more recent studies have examined cost-sensitive reinforcement-learning approaches for credit-risk decisions (Herasymovych et al., 2019; C-Rella et al., 2025).

Despite this potential, the use of reinforcement learning in wholesale credit decisioning raises important concerns. Reinforcement-learning policies may be difficult to interpret because recommendations are influenced by historical trajectories, reward design, estimated future outcomes, and interactions among numerous risk variables. In banking, a recommendation cannot be accepted merely because it improves a technical performance metric. Credit officers, risk managers, validators, auditors, and regulators need to understand why a particular action was recommended, which risk factors drove the decision, whether the recommendation complies with policy limits, and whether an alternative action would have been appropriate under different conditions.

Explainable artificial intelligence offers methods for making model outputs more understandable. In credit-risk applications, feature attribution, local explanations, counterfactual analysis, and rule-based representations can help users identify the factors associated with a prediction or decision (Gramegna & Giudici, 2021; Dastile & Çelik, 2021). However, explainability in reinforcement learning requires more than identifying influential features. It must also explain the relationship between the observed credit state, the selected action, the reward structure, and the expected future effect of alternative actions. Explainable reinforcement learning therefore provides a relevant foundation for developing decision-support systems that are adaptive while remaining transparent and accountable (Puiutta & Veith, 2020; Milani et al., 2024).

This study proposes an explainable reinforcement-learning framework for real-time credit-risk decisioning in wholesale banking ecosystems. The framework is intended to support sequential credit decisions using borrower, facility, collateral, behavioural, portfolio, and market data. It combines offline reinforcement learning with policy constraints, explanation mechanisms, and human oversight procedures. The study addresses four questions: how wholesale credit decisions can be represented as a sequential learning problem; how explanations can be generated for recommended actions; whether the proposed approach can improve risk-adjusted decision quality compared with conventional models; and how governance controls can support safe adoption.

The study makes three main contributions. First, it develops a structured framework that links wholesale credit-risk data, sequential decision policies, explainability outputs, and governance requirements. Second, it proposes an evaluation approach that extends beyond predictive accuracy to include expected-loss control, risk-adjusted return, policy compliance, decision latency, robustness, and explanation quality. Third, it provides practical guidance for adopting reinforcement learning in a high-impact financial environment where autonomous decisioning must remain constrained by credit policy, audit requirements, and human judgement.

II. LITERATURE REVIEW

1. Credit Risk Decisioning in Wholesale Banking

Credit-risk decisioning in wholesale banking involves the assessment and ongoing management of financial exposures to corporations, institutions, sovereign entities, and other large counterparties. Decisions commonly cover facility approval, credit limits, collateral arrangements, covenant requirements, pricing, restructuring, and monitoring intensity. These decisions are usually based on financial statements, cash-flow performance, repayment behaviour, internal ratings, collateral quality, industry conditions, market indicators, and portfolio concentration.

The complexity of wholesale credit management arises from the size and structure of exposures. A single borrower may hold multiple facilities, guarantees, trade-finance instruments, derivatives, or committed credit lines. Changes in one facility may influence the risk profile of other exposures within the same relationship. In addition, borrower conditions can deteriorate quickly when macroeconomic, sectoral, or market conditions change. For this reason, periodic review processes alone may be insufficient for identifying early warning signals and supporting timely intervention.

Traditional credit-risk management relies heavily on rating systems, scorecards, expert judgement, and rule-based approval processes. These approaches provide consistency and support regulatory reporting, but they may not adapt easily to complex and changing patterns in borrower behaviour. Machine-learning methods have therefore gained attention because they can identify nonlinear associations and interactions among large volumes of structured and unstructured data. Systematic reviews show that machine learning can improve credit-risk prediction when compared with some conventional statistical approaches, particularly where datasets include diverse financial and behavioural variables (Shi et al., 2022; Noriega et al., 2023).

However, prediction does not directly solve the decision problem. A model that identifies a high probability of borrower distress does not necessarily determine whether the bank should reduce a limit, reprice a facility, request additional collateral, intensify monitoring, or escalate the account. These actions have different economic, operational, and relationship consequences. Therefore, wholesale credit decisioning requires approaches that can balance current risk indicators with expected future outcomes.

2. Machine Learning and Explainability in Credit Risk

Machine-learning models are increasingly used in credit-risk assessment because they can process complex relationships among borrower characteristics, financial indicators, behavioural signals, and macroeconomic variables. Random forests, gradient-boosting models, support vector machines, neural networks, and ensemble methods have been applied to credit scoring, default prediction, early-warning systems, and loss estimation. These methods can improve discrimination and predictive accuracy, especially where risk relationships are nonlinear or where variables interact in ways that are difficult to capture with traditional regression models.

Despite these advantages, model complexity may create challenges for financial institutions. Credit decisions affect access to funding, pricing, collateral obligations, and long-term commercial relationships. As a result, institutions must be able to justify model-supported recommendations to internal reviewers, customers where appropriate, auditors, and regulatory authorities. Bussmann et al. (2020) argue that explainability is particularly important in financial technology because models should support transparent decision processes rather than operate as unexamined technical systems.

Explainable AI methods aim to provide understandable evidence for model outputs. In credit risk, commonly used approaches include SHAP values, LIME, feature-importance rankings, partial-dependence analysis, local surrogate models, and counterfactual explanations. SHAP-based methods can identify the relative contribution of variables to an individual prediction, while counterfactual explanations can indicate which changes in input conditions may lead to a different outcome. Gramegna and Giudici (2021) found that explanation approaches should be evaluated carefully because they may differ in discriminatory power and stability when applied to credit-risk models.

Explainability is also linked to fairness and responsible use. Machine-learning models may reproduce historical bias if their training data reflects unequal lending patterns, inconsistent approval practices, or incomplete borrower information. Fuster et al. (2022) show that machine learning can affect the distribution of credit access in ways that may not be evenly shared across borrowers. Therefore, explainable credit-risk systems should be supported by bias testing, governance controls, data-quality assurance, and expert review.

3. Reinforcement Learning for Sequential Financial Decisions

Reinforcement learning is a decision-learning method in which an agent observes a state, selects an action, receives a reward, and updates its policy based on the resulting outcomes. Unlike supervised learning, which usually predicts a target value from historical examples, reinforcement learning aims to learn a policy that maximises cumulative reward over a sequence of decisions. This makes it suitable for situations where actions influence future states and where the best decision depends on longer-term outcomes.

In a wholesale credit setting, the state may include borrower financial health, internal rating, payment behaviour, covenant status, collateral coverage, facility utilisation, sector outlook, and portfolio concentration. Possible actions may include approving a facility, maintaining an exposure limit, reducing a limit, revising pricing, requesting collateral, placing an account under enhanced monitoring, or escalating the case for human review. The reward may incorporate expected loss, risk-adjusted return, capital use, policy compliance, and decision timeliness.

Herasymovych et al. (2019) demonstrated the potential of reinforcement learning for optimising acceptance thresholds in credit scoring. Their study showed that a reinforcement-learning approach can improve decision policies by considering long-term business outcomes rather than relying only on a fixed threshold. More recently, C-Rella et al. (2025) examined cost-sensitive reinforcement learning for credit risk, highlighting the importance of aligning rewards with the asymmetric cost of credit losses and incorrect lending decisions.

Although reinforcement learning offers benefits for dynamic decisioning, its use in banking must be approached cautiously. Financial institutions cannot safely allow models to learn through unrestricted live experimentation where decisions may affect high-value credit exposures. Offline reinforcement learning is therefore more suitable because it learns from historical datasets rather than interacting directly with live credit portfolios during training. Conservative Q-learning, implicit Q-learning, and related approaches aim to reduce the risk of selecting actions that are poorly represented in historical data (Kumar et al., 2020; Kostrikov et al., 2022).

Offline reinforcement learning also requires careful policy evaluation. A new decision policy cannot be assumed to perform well simply because it achieves high estimated reward during training. Off-policy evaluation methods are needed to estimate how a proposed policy may perform using historical trajectories. Kallus and Uehara (2020) and Voloshin et al. (2021) highlight the importance of robust off-policy evaluation because reinforcement-learning policies may be sensitive to data coverage, transition assumptions, and model misspecification.

4. Explainable Reinforcement Learning

Explainable reinforcement learning extends conventional reinforcement learning by providing information about why an agent selected an action, how its expected return was formed, and what alternative actions were considered. This is important because reinforcement-learning recommendations are not based only on individual feature values. They depend on the policy's estimate of future outcomes, reward structure, state transitions, and constraints.

Puiutta and Veith (2020) categorise explainable reinforcement learning into approaches that explain policies, actions, rewards, states, and environment dynamics. Policy explanations clarify how the agent tends to act under different conditions. Action explanations provide reasons for a specific recommendation in a particular case. Reward explanations show how competing objectives influenced the action, while state explanations identify the most relevant inputs used by the policy. These explanation types are particularly relevant for wholesale credit decisions because users need to understand not only which factors increased risk, but also why one intervention was preferred over another.

Madumal et al. (2020) propose a causal perspective for explainable reinforcement learning, showing that explanations can be structured around cause-and-effect relationships between state conditions, actions, and outcomes. This perspective is useful for credit-risk decisioning because it can support counterfactual analysis. For example, a system may indicate that a facility-limit reduction was recommended because of declining debt-service capacity, repeated covenant breaches, and falling collateral coverage. It may also show that maintaining the existing limit could have been considered if collateral coverage had remained above a defined threshold and cash-flow volatility had decreased.

Recent surveys have further emphasised the need for explanations that are accurate, stable, comprehensible, and useful to different users. Milani et al. (2024) note that explainable reinforcement learning should not be assessed only by whether an explanation is available. It should also be assessed by whether the explanation faithfully represents the policy, remains consistent under small input changes, and supports appropriate human judgement. In banking, these requirements are essential because explanations must serve credit officers, model validators, internal auditors, and senior risk managers.

5. Research Gap

The literature shows growing interest in machine learning, explainable AI, and reinforcement learning for financial decision-making. However, several gaps remain. First, much of the credit-risk literature focuses on one-time prediction tasks such as default classification, rating estimation, or credit-score optimisation. There is less attention to sequential wholesale credit decisions involving exposure management, collateral actions, pricing adjustments, monitoring, and escalation.

Second, existing reinforcement-learning studies in credit risk often focus on consumer credit or threshold optimisation rather than the broader ecosystem of wholesale banking. Wholesale credit portfolios require consideration of interconnected facilities, large exposures, portfolio concentration, covenant events, collateral movements, and relationship-level risk.

Third, explainability studies in credit risk often concentrate on supervised models. While these methods are useful, they do not fully address the need to explain long-term policy decisions made by reinforcement-learning agents. There is limited research on combining reward decomposition, policy rationale, counterfactual analysis, and governance controls within a single framework for wholesale credit-risk decisioning.

This study addresses these gaps by proposing an explainable reinforcement-learning framework designed for real-time credit-risk decisioning in wholesale banking. It treats credit management as a sequential process, incorporates policy constraints and human oversight, and evaluates the approach using financial, operational, governance, and explanation-quality measures.

Table 1. Comparative Review of Existing Approaches to Credit-Risk Decisioning

Approach	Main Strength	Main Limitation	Suitability for Real-Time Whole sale Credit Decisions
Logistic regression and scorecards	Transparent and stable	Limited nonlinear and sequential learning capability	Moderate
Tree-based machine learning	Strong predictive accuracy	May not optimize multi-period decisions	Moderate to high
Deep learning models	Captures complex relations hips	Low interpretability	Moderate
Reinforcement learning	Optimizes sequential decisions	Requires careful reward design and safety controls	High
Explainable reinforcement learning	Supports adaptive and interpret able policies	Methodological Complexity and validation burden	High

III. CONCEPTUAL FOUNDATION AND RESEARCH FRAMEWORK

1. Conceptual Foundation

The proposed framework is grounded in sequential decision theory, credit-risk management, explainable artificial intelligence, and model governance. It treats wholesale credit management as a sequence of interconnected decisions rather than a one-time classification task. A bank must continuously assess whether an existing or proposed credit exposure remains consistent with the borrower's financial condition, collateral position, facility use, sector outlook, and the bank's risk appetite.

Reinforcement learning is suitable for this setting because it learns a decision policy from repeated state-action-outcome relationships. At each decision point, the agent observes the current credit state, selects an action, and receives a reward based on the resulting financial and risk outcomes. The objective is not simply to predict default, but to identify the action that is expected to produce the most appropriate long-term result. In wholesale banking, these actions may include maintaining a facility limit, reducing exposure, revising risk-based pricing, requesting additional collateral, initiating enhanced monitoring, or referring the case to a credit committee.

The framework also recognises that banking decisions cannot be based solely on automated optimisation. Credit decisions affect borrowers, counterparties, revenue, capital consumption, and portfolio concentration. Therefore, the reinforcement-learning policy must operate within approved credit-policy limits and remain subject to human review where required. Offline reinforcement learning is appropriate because the policy can be trained from historical credit trajectories without exposing live portfolios to experimental actions. Conservative approaches are particularly relevant where historical data does not adequately represent all possible decision paths (Kumar et al., 2020; Prudencio et al., 2024).

Explainability is a central component of the framework. A credit officer needs to understand not only the recommended action, but also the reasons behind it. The system must identify the main borrower, facility, collateral, and market conditions that influenced the recommendation. It must also show how the expected reward was formed and why the selected action was preferred to alternative actions. Explainable reinforcement learning supports this requirement by linking state conditions, reward components, policy choices, and estimated future outcomes (Puiutta & Veith, 2020; Madumal et al., 2020).

2. Wholesale Banking Ecosystem Model

The framework views wholesale credit decisioning as an ecosystem involving several internal and external data sources. Internal sources include borrower financial statements, internal ratings, repayment records, facility utilisation, collateral values, covenant status, exposure limits, and previous credit decisions. External sources may include market indicators, sector performance data, macroeconomic conditions, legal events, and counterparty information.

The main users of the framework are relationship managers, credit analysts, credit approvers, risk managers, model validators, and internal auditors. Each user requires a different level of explanation. Relationship managers may need a concise account of key borrower risks and recommended actions. Credit officers may require detailed evidence concerning financial deterioration, collateral adequacy, covenant performance, and risk-adjusted return. Model validators and auditors require traceable records of data inputs, model versions, policy constraints, explanation outputs, and human overrides.

The proposed framework therefore combines automated recommendation with formal governance processes. It does not replace credit judgement. Instead, it provides a structured basis for identifying risk changes, evaluating response options, and supporting timely escalation.

3. Reinforcement-Learning Decision Model
The reinforcement-learning environment is defined through four main elements: state, action, reward, and transition.

The state represents the current risk condition of a borrower and facility. It may include liquidity, profitability, leverage, repayment behaviour, collateral coverage, covenant compliance, exposure concentration, sector outlook, market volatility, and internal credit rating. The state is updated as new information becomes available.

The action represents the credit response selected by the model. Actions may include approving a request, maintaining an existing limit, reducing exposure, increasing pricing, requesting additional collateral, initiating enhanced monitoring, or escalating the case for manual approval.

The reward reflects the value of the selected action. It should balance risk-adjusted return, expected loss, capital cost, policy compliance, and decision timeliness. A suitable reward function can be represented as:

$$R_t = \alpha(RAROC_t) - \beta(EL_t) - \gamma(CC_t) \\ - \delta(PV_t) - \lambda(DT_t)$$

Where:
- $RAROC_t$ represents risk-adjusted return on capital at time t.
- EL_t represents expected loss.
- CC_t represents capital cost.
- PV_t represents policy-violation penalties.
- DT_t represents decision delay.
- α, β, γ, δ, and λ represent weights determined by the institution's risk appetite and operating priorities.

The transition reflects how borrower and facility conditions change after an action is taken. For example, a request for additional collateral may improve loss protection, while delayed intervention may increase expected loss if borrower performance deteriorates. The model learns from historical sequences of these transitions to estimate the likely long-term effect of each action.

4. Explainability Layer
The explainability layer provides four forms of explanation for each recommendation.

First, state explanations identify the variables that most influenced the decision. These may include falling liquidity, declining cash-flow coverage, repeated covenant breaches, declining collateral value, increased utilisation, or adverse sector conditions.

Second, reward explanations show how the main components of the reward function affected the recommendation. For example, an action may be preferred because it reduces expected loss and improves capital efficiency, even where it lowers immediate revenue.

Third, policy explanations describe why the selected action was considered more appropriate than the available alternatives. This is important where the model recommends escalation rather than automatic approval or decline.

Fourth, counterfactual explanations show what changes in borrower or facility conditions could have resulted in a different action. For example, the model may indicate that maintaining an exposure limit would have been more appropriate if collateral coverage had remained above a required threshold and covenant compliance had improved.

This approach extends conventional feature-attribution methods by explaining the decision path rather than only identifying important variables. Such explanations are important for high-impact financial decisions, where users must assess whether a recommendation is reasonable, policy compliant, and suitable for implementation (Milani et al., 2024).

5. Proposed Research Framework
The proposed research framework consists of five connected layers:

International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 4, July-Aug-2026, ISSN (Online): 2395-566X

- Data Integration Layer: Combines borrower, facility, collateral, transaction, covenant, portfolio, and market data.
- Credit-Risk State Construction Layer: Converts the integrated data into dynamic credit-risk states.
- Explainable Reinforcement-Learning Layer: Learns a policy that recommends appropriate credit actions from historical decision trajectories.
- Human Oversight and Policy-Control Layer: Applies approval limits, risk-appetite rules, exception controls, and manual review requirements.
- Decision Execution and Monitoring Layer: Records recommendations, actions, outcomes, overrides, explanation outputs, and model-performance indicators.

The framework includes feedback from decision outcomes to the data and learning layers. This allows the model to be reviewed and recalibrated as borrower behaviour, portfolio conditions, and market environments change.

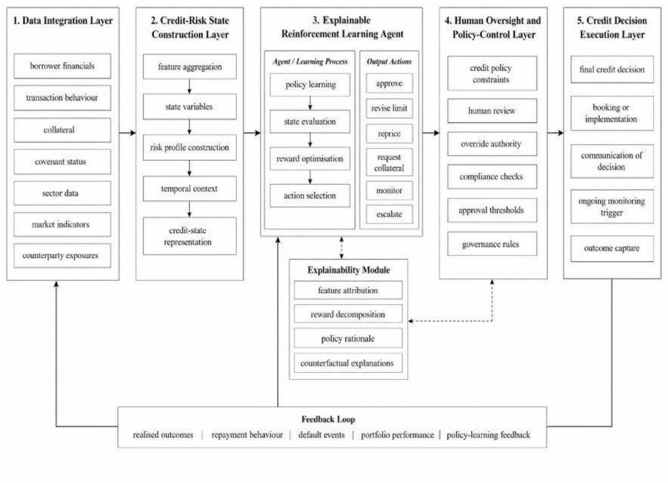

Fig 1. Explainable reinforcement-learning architecture for controlled credit-risk decision-making.

IV. RESEARCH METHODOLOGY

1. Research Design

This study adopts a design science and empirical evaluation approach. The design science component focuses on the development of an explainable reinforcement-learning framework for wholesale credit-risk decisioning. The empirical component evaluates whether the proposed framework improves decision quality when compared with conventional credit-risk models and non-explainable reinforcement-learning approaches.

The research process consists of five stages: data preparation, environment design, model development, explanation generation, and comparative evaluation. This structure supports both the technical development of the framework and the assessment of its practical relevance for regulated banking environments.

2. Data Sources and Unit of Analysis

The study uses historical wholesale credit data collected from banking systems or a suitably representative institutional dataset. Relevant data sources may include borrower financial statements, internal credit ratings, repayment records, collateral registers, covenant-monitoring records, facility utilisation, exposure limits, restructuring events, defaults, recoveries, and external market indicators.

The unit of analysis is the borrower-facility-period observation. Each observation represents the risk condition of a borrower and a specific credit facility at a defined time interval, such as monthly, quarterly, or following a material credit event. This approach allows the study to capture how borrower conditions and credit decisions evolve over time.

To reduce information leakage, the data should be arranged chronologically. Variables used to predict or recommend an action must be available at the time the decision is assumed to occur. Future outcomes, such as realised default or recovery, should only be used to calculate rewards and evaluate the policy.

3. Data Preparation and Feature Construction

Data preparation begins with the integration of information from credit, finance, collateral, transaction, and risk-management systems. Duplicate records, missing values, inconsistent borrower identifiers, and timing differences across systems should be addressed before model development.

The dataset should include four categories of variables:

- Borrower variables: Liquidity ratios, profitability, debt-service capacity, leverage, cash-flow stability, and internal rating.
- Facility variables: Exposure amount, maturity, utilisation, pricing, collateral coverage, guarantee status, and covenant requirements.
- Behavioural variables: Repayment delays, covenant breaches, limit excesses, restructuring requests, and changes in account activity.

- External variables: Interest rates, sector conditions, commodity prices, foreign-exchange movements, and macroeconomic indicators.

Continuous variables should be normalised where necessary, while categorical variables should be encoded using appropriate methods. Outliers should be reviewed carefully because extreme values may reflect genuine borrower distress rather than data errors. Missing values should be handled using documented imputation procedures, with missingness indicators retained where the absence of information may itself signal risk.

4. Reinforcement-Learning Environment Design

The learning environment represents the sequence of credit decisions made over time. The state includes the borrower's current risk profile, while the action represents the selected credit response. The transition reflects how borrower and facility conditions evolve after the action, and the reward captures the financial and risk consequences of that action.

Historical decision records provide the behaviour policy from which the reinforcement-learning agent learns. Because the agent is trained from past data, the study should use offline reinforcement-learning methods rather than live exploration. Conservative Q-learning or implicit Q-learning may be applied because these methods are designed to reduce overestimation of actions that are poorly represented in historical data (Kumar et al., 2020; Kostrikov et al., 2022).

The action space should be limited to actions that are realistic within wholesale banking practice. Actions may include maintaining a facility, reducing exposure, increasing pricing, requesting collateral, escalating for review, or placing the borrower under enhanced monitoring. Actions that violate policy limits or approval authority should be excluded through predefined constraints.

5. Model Development

The study should develop at least one explainable reinforcement-learning model and compare it with several baseline approaches. Suggested baseline models include logistic regression, gradient boosting, a deep-learning classifier, a rule-based credit-policy model, and a conventional reinforcement-learning model without the explanation layer.
The model development process should include chronological training, validation, and test datasets. A suitable allocation may be based on time periods rather than random sampling. This allows the model to be evaluated under conditions that resemble real credit decisioning, where decisions are made using past information and assessed against future outcomes.

Hyperparameters should be selected using the validation set. These may include learning rate, discount factor, batch size, target-network update frequency, and reward weights. The final test set should remain unseen until the model configuration has been completed.

6. Explainability Procedure

The proposed model should produce explanations alongside each action recommendation. The explanation procedure should include feature attribution, reward decomposition, policy comparison, and counterfactual analysis.

Feature attribution identifies the borrower and facility conditions that contributed most strongly to the recommendation. Reward decomposition explains how expected loss, risk-adjusted return, capital cost, policy compliance, and decision timeliness influenced the action. Policy comparison identifies why the chosen action was preferred over competing actions. Counterfactual analysis identifies changes in input conditions that could have resulted in a different action.

Explanation quality should be assessed using fidelity, stability, sparsity, and expert usefulness. Fidelity measures whether the explanation accurately represents the model's actual reasoning. Stability assesses whether similar credit states produce similar explanations. Sparsity considers whether explanations focus on a manageable number of important factors. Expert usefulness can be evaluated through structured assessment by credit analysts or risk officers.

7. Evaluation Strategy

The framework should be evaluated across five dimensions.
First, credit-risk performance should be assessed using measures such as AUC-ROC, precision, recall, F1-score, calibration error, and expected-loss estimation quality.

Second, decision performance should be assessed through risk-adjusted return, expected-loss reduction, capital efficiency, and the quality of recommended actions.

Third, operational performance should include decision latency, escalation rate, manual override rate, and processing efficiency.

Fourth, governance performance should assess policy-compliance rate, constraint violations, audit-log completeness, explanation coverage, and reproducibility.

Fifth, explanation quality should assess fidelity, stability, comprehensibility, and usefulness to expert users.

The model should also be tested under stress conditions. These may include borrower deterioration, sector downturns, collateral-value decline, higher interest rates, and incomplete data. Such tests help determine whether the policy remains appropriate when credit conditions become adverse.

8. Offline Policy Evaluation and Validation

A key methodological requirement is the evaluation of the learned policy without applying it directly to live credit decisions. Off-policy evaluation methods should therefore be used to estimate how the proposed policy would have performed using historical credit trajectories. Suitable approaches may include importance sampling, doubly robust estimation, and model-based simulation.

Kallus and Uehara (2020) emphasise that off-policy evaluation is necessary when assessing policies that differ from historical decision practices. Voloshin et al. (2021) similarly show that policy-evaluation results may vary substantially depending on data coverage and estimator choice. For this reason, the study should apply more than one evaluation method and compare the consistency of results.

9. Ethical and Governance Considerations

The study should protect borrower confidentiality and comply with relevant data-protection requirements. Personally identifiable information should be removed or anonymised where possible. Access to credit data should be restricted to authorised users, and model outputs should be logged for audit purposes.

The framework should also include fairness and bias assessment. Historical credit decisions may reflect past inconsistencies or unequal treatment. Therefore, the model should be tested for systematic differences in recommendations across relevant borrower groups, sectors, or regions where such testing is legally and operationally appropriate. Human oversight remains necessary for high-value, unusual, or low-confidence cases.

IV. RESEARCH METHODOLOGY

1. Research Design

This study adopts a design science and empirical evaluation approach. The design science component focuses on the development of an explainable reinforcement-learning framework for wholesale credit-risk decisioning. The empirical component evaluates whether the proposed framework improves decision quality when compared with conventional credit-risk models and non-explainable reinforcement-learning approaches.

The research process consists of five stages: data preparation, environment design, model development, explanation generation, and comparative evaluation. This structure supports both the technical development of the framework and the assessment of its practical relevance for regulated banking environments.

2. Data Sources and Unit of Analysis

The study uses historical wholesale credit data collected from banking systems or a suitably representative institutional dataset. Relevant data sources may include borrower financial statements, internal credit ratings, repayment records, collateral registers, covenant-monitoring records, facility utilisation, exposure limits, restructuring events, defaults, recoveries, and external market indicators.

The unit of analysis is the borrower-facility-period observation. Each observation represents the risk condition of a borrower and a specific credit facility at a defined time interval, such as monthly, quarterly, or following a material credit event. This approach allows the study to capture how borrower conditions and credit decisions evolve over time.

To reduce information leakage, the data should be arranged chronologically. Variables used to predict or recommend an action must be available at the time the decision is assumed to occur. Future outcomes, such as realised default or recovery, should only be used to calculate rewards and evaluate the policy.

3. Data Preparation and Feature Construction

Data preparation begins with the integration of information from credit, finance, collateral, transaction, and risk-management systems. Duplicate records, missing values, inconsistent borrower identifiers, and timing differences across systems should be addressed before model development.

The dataset should include four categories of variables:

- Borrower variables: Liquidity ratios, profitability, debt-service capacity, leverage, cash-flow stability, and internal rating.
- Facility variables: Exposure amount, maturity, utilisation, pricing, collateral coverage, guarantee status, and covenant requirements.
- Behavioural variables: Repayment delays, covenant breaches, limit excesses, restructuring requests, and changes in account activity.
- External variables: Interest rates, sector conditions, commodity prices, foreign-exchange movements, and macroeconomic indicators.

Continuous variables should be normalised where necessary, while categorical variables should be encoded using appropriate methods. Outliers should be reviewed carefully because extreme values may reflect genuine borrower distress rather than data errors. Missing values should be handled using documented imputation procedures, with missingness indicators retained where the absence of information may itself signal risk.

4. Reinforcement-Learning Environment Design

The learning environment represents the sequence of credit decisions made over time. The state includes the borrower's current risk profile, while the action represents the selected credit response. The transition reflects how borrower and facility conditions evolve after the action, and the reward captures the financial and risk consequences of that action.

Historical decision records provide the behaviour policy from which the reinforcement-learning agent learns. Because the agent is trained from past data, the study should use offline reinforcement-learning methods rather than live exploration. Conservative Q-learning or implicit Q-learning may be applied because these methods are designed to reduce overestimation of actions that are poorly represented in historical data (Kumar et al., 2020; Kostrikov et al., 2022).

The action space should be limited to actions that are realistic within wholesale banking practice. Actions may include maintaining a facility, reducing exposure, increasing pricing, requesting collateral, escalating for review, or placing the borrower under enhanced monitoring. Actions that violate policy limits or approval authority should be excluded through predefined constraints.

5. Model Development

The study should develop at least one explainable reinforcement-learning model and compare it with several baseline approaches. Suggested baseline models include logistic regression, gradient boosting, a deep-learning classifier, a rule-based credit-policy model, and a conventional reinforcement-learning model without the explanation layer.

The model development process should include chronological training, validation, and test datasets. A suitable allocation may be based on time periods rather than random sampling. This allows the model to be evaluated under conditions that resemble real credit decisioning, where decisions are made using past information and assessed against future outcomes.
Hyperparameters should be selected using the validation set. These may include learning rate, discount factor, batch size, target-network update frequency, and reward weights. The final test set should remain unseen until the model configuration has been completed.

6. Explainability Procedure

The proposed model should produce explanations alongside each action recommendation. The explanation procedure should include feature attribution, reward decomposition, policy comparison, and counterfactual analysis.

Feature attribution identifies the borrower and facility conditions that contributed most strongly to the recommendation. Reward decomposition explains how expected loss, risk-adjusted return, capital cost, policy compliance, and decision timeliness influenced the action. Policy comparison identifies why the chosen action was preferred over competing actions. Counterfactual analysis identifies changes in input conditions that could have resulted in a different action.

Explanation quality should be assessed using fidelity, stability, sparsity, and expert usefulness. Fidelity measures whether the explanation accurately represents the model's actual reasoning. Stability assesses whether similar credit states produce similar explanations. Sparsity considers whether explanations focus on a manageable number of important factors. Expert usefulness can be evaluated through structured assessment by credit analysts or risk officers.

7. Evaluation Strategy

The framework should be evaluated across five dimensions.
First, credit-risk performance should be assessed using measures such as AUC-ROC, precision, recall, F1-score, calibration error, and expected-loss estimation quality.

Second, decision performance should be assessed through risk-adjusted return, expected-loss reduction, capital efficiency, and the quality of recommended actions.

Third, operational performance should include decision latency, escalation rate, manual override rate, and processing efficiency.

Fourth, governance performance should assess policy-compliance rate, constraint violations, audit-log completeness, explanation coverage, and reproducibility.

Fifth, explanation quality should assess fidelity, stability, comprehensibility, and usefulness to expert users.

The model should also be tested under stress conditions. These may include borrower deterioration, sector downturns, collateral-value decline, higher interest rates, and incomplete data. Such tests help determine whether the policy remains appropriate when credit conditions become adverse.

8. Offline Policy Evaluation and Validation
A key methodological requirement is the evaluation of the learned policy without applying it directly to live credit decisions. Off-policy evaluation methods should therefore be used to estimate how the proposed policy would have performed using historical credit trajectories. Suitable approaches may include importance sampling, doubly robust estimation, and model-based simulation.

Kallus and Uehara (2020) emphasise that off-policy evaluation is necessary when assessing policies that differ from historical decision practices. Voloshin et al. (2021) similarly show that policy-evaluation results may vary substantially depending on data coverage and estimator choice. For this reason, the study should apply more than one evaluation method and compare the consistency of results.

9. Ethical and Governance Considerations
The study should protect borrower confidentiality and comply with relevant data-protection requirements. Personally identifiable information should be removed or anonymised where possible. Access to credit data should be restricted to authorised users, and model outputs should be logged for audit purposes.

The framework should also include fairness and bias assessment. Historical credit decisions may reflect past inconsistencies or unequal treatment. Therefore, the model should be tested for systematic differences in recommendations across relevant borrower groups, sectors, or regions where such testing is legally and operationally appropriate. Human oversight remains necessary for high-value, unusual, or low-confidence cases.

V. PROPOSED EXPLAINABLE REINFORCEMENT LEARNING FRAMEWORK

1. Framework Overview
The proposed framework supports real-time credit-risk decisioning across wholesale banking activities. It combines integrated risk data, offline reinforcement learning, explanation mechanisms, policy constraints, and human oversight. Its purpose is to assist credit professionals in selecting actions that balance expected loss, risk-adjusted return, capital use, timeliness, and compliance with approved credit policies.

The framework is designed as a decision-support system rather than a fully autonomous credit approval mechanism. It provides recommendations for relationship managers, credit analysts, credit approvers, and risk managers, while preserving formal approval structures for material or exceptional decisions. This approach is appropriate because wholesale credit decisions often involve large and concentrated exposures, complex borrower structures, and significant portfolio implications.

The framework contains six connected components: data integration, credit-state construction, reinforcement-learning policy generation, policy-control mechanisms, explainability outputs, and continuous monitoring.

2. Data Integration and Credit-State Construction
The data integration component brings together information from credit-risk, finance, collateral, transaction, treasury, and portfolio-management systems. Relevant inputs include borrower financial statements, internal ratings, repayment patterns, covenant compliance, facility utilisation, collateral values, exposure limits, industry conditions, market indicators, and counterparty relationships.

These inputs are transformed into a dynamic credit-risk state. The state represents the current financial and risk position of a borrower and its related facilities at a particular decision point. It may include liquidity, profitability, leverage, debt-service capacity, payment behaviour, collateral coverage, exposure concentration, sector risk, and macroeconomic conditions.

The construction of the state must follow the actual sequence of information availability. Data recorded after a credit decision should not be used as an input for that earlier decision. This reduces information leakage and ensures that the model reflects the conditions available to the bank at the time of recommendation.

3. Reinforcement-Learning Policy Engine
The reinforcement-learning policy engine receives the current credit-risk state and recommends an action from a controlled set of credit responses. The action space should reflect established wholesale banking practice. Possible actions include approving a facility, maintaining an existing limit, reducing exposure, revising risk-based pricing, requesting additional collateral, initiating enhanced monitoring, escalating a case for manual review, or beginning a restructuring assessment.

The model is trained through offline reinforcement learning using historical credit-decision records. This approach is preferable to direct live learning because a bank cannot safely test experimental actions on active high-value exposures. Offline methods learn from previous borrower trajectories, historical decisions, and realised outcomes without introducing unapproved risk into the live portfolio.

Conservative Q-learning can be used to reduce overestimation of actions that are poorly represented in historical data. This is important where past credit decisions provide limited evidence about unusual or rarely applied actions. Implicit Q-learning may also be suitable for learning from complex historical decision patterns while remaining within the available data distribution (Kumar et al., 2020; Kostrikov et al., 2022).

4. Reward Design

The reward function directs the model towards actions that support sustainable credit outcomes. A suitable reward function should not focus only on short-term revenue because this could encourage actions that increase future credit losses. Instead, it should incorporate risk-adjusted return, expected loss, capital cost, policy compliance, and decision timeliness.

$$R_t = \alpha(RAROC_t) - \beta(EL_t) - \gamma(CC_t) \\ - \delta(PV_t) - \lambda(DT_t)$$

Where:

- $RAROCt$represents risk-adjusted return on capital at time t.
- ELtrepresents expected loss.
- CCtrepresents capital cost.
- PVtrepresents penalties for policy violations.
- DTtrepresents decision delay.
- α, β, γ, δ, and $\lambda$are weights determined by the bank's risk appetite and credit-management priorities.

This formulation allows the model to recognise that a lower-revenue action may still be preferable where it reduces expected loss, protects collateral coverage, avoids a policy breach, or limits future deterioration in borrower quality. Cost-sensitive reinforcement learning is especially relevant in credit-risk settings because incorrect approval and incorrect restriction decisions may have materially different financial consequences (C-Rella et al., 2025).

5. Explainability Module

The explainability module provides supporting evidence for every recommended action. It is intended to help users understand the recommendation, review its consistency with policy, and assess whether a human intervention is required.

The first output is a state explanation. This identifies the borrower, facility, collateral, portfolio, and market variables that had the strongest influence on the recommendation. For example, the model may identify declining liquidity, repeated covenant breaches, increased facility utilisation, falling collateral values, or adverse sector conditions as major decision drivers.

The second output is a reward explanation. This shows how expected loss, risk-adjusted return, capital use, policy compliance, and decision delay contributed to the chosen action. A recommendation to reduce a facility limit may therefore be explained by a substantial expected-loss reduction and lower capital consumption, despite lower short-term revenue.

The third output is a policy explanation. This compares the selected action with other available actions. It helps users understand why enhanced monitoring may have been preferred to maintaining the current limit, or why escalation was selected instead of immediate exposure reduction.

The fourth output is a counterfactual explanation. This identifies the changes in borrower or facility conditions that could have led to a different recommendation. For instance, the model may indicate that maintaining an existing limit would have been appropriate if debt-service coverage had improved, collateral coverage had remained adequate, and covenant compliance had stabilised.

These explanation types are consistent with the explainable reinforcement-learning literature, which distinguishes between explanations of states, actions, policies, rewards, and causal outcomes (Puiutta & Veith, 2020; Madumal et al., 2020; Milani et al., 2024).

6. Human Oversight and Policy Controls

The framework includes a human oversight layer because certain decisions require formal judgement, delegated approval, or committee review. Cases involving high-value exposures, significant covenant breaches, incomplete data, unusual borrower structures, or low-confidence recommendations should be automatically escalated for manual assessment.

The model must also operate within predefined policy constraints. It should not recommend actions that breach approved exposure limits, collateral requirements, sector concentration thresholds, internal rating rules, or delegated lending authority. The constraint layer therefore acts as a safeguard between the learned policy and operational execution.

Human reviewers should be able to accept, modify, or reject model recommendations. Every override should be documented with the reason for the change, the responsible decision-maker, and the supporting evidence. These records can reveal recurring weaknesses in the model, changes in credit policy, or important patterns that were not fully represented in the historical training data.

Table 3. Explainability Outputs for Wholesale Credit-Risk Users

User Group	Main Requirement	Explanation Output
Relationship manager	Understanding of borrower risk changes	Key risk Drivers and recommended action
Credit analyst	Detailed credit assessment	Financial, behavioural, collateral, and covenant evidence
Credit approver	Basis for approval, rejection, or escalation	Reward trade-off, alternative actions, and policy constraints
Risk manager	Portfolio and risk-appetite impact	Exposure effect, expected-loss effect, and concentration implications
Model validator	Evidence of model reliability	Policy logic, reward design, stability, and performance history
Internal auditor	Decision traceability	Data lineage, model version, explanation record, and override history

7. Monitoring and Feedback

The framework should continuously monitor recommendation outcomes, borrower performance, collateral movements, exposure changes, policy exceptions, and override patterns. These results should be used to identify model drift, declining data quality, changing portfolio conditions, and weaknesses in the reward structure.

Periodic recalibration is necessary because borrower behaviour, economic conditions, market volatility, and credit policy may change over time. The monitoring process should also assess whether the explanations remain stable, relevant, and understandable for credit users. Explainability must remain part of model review rather than being treated as a one-time implementation feature.

VI. EXPERIMENTAL DESIGN AND EVALUATION FRAMEWORK

1. Comparative Model Design

The proposed explainable reinforcement-learning model should be evaluated against conventional credit-risk approaches and advanced machine-learning alternatives. This comparison is necessary to determine whether the framework improves credit decision quality without reducing transparency, policy compliance, or operational reliability.

The baseline models should include logistic regression, gradient boosting, a deep-learning credit-risk classifier, a rule-based credit-policy engine, and a non-explainable reinforcement-learning model. Logistic regression provides a transparent benchmark, while gradient boosting and deep learning represent stronger nonlinear prediction approaches. The rule-based model represents conventional policy-led decisioning, while the non-explainable reinforcement-learning model allows the study to assess the practical contribution of the explanation component.

2. Dataset Partitioning and Validation

The dataset should be divided chronologically into training, validation, and test periods. This approach is preferable to random sampling because it reflects how credit decisions occur in practice. The model should learn from earlier borrower and facility observations, then be tested on later periods where economic conditions and risk patterns may have changed.

The training dataset should be used to estimate the policy parameters. The validation dataset should support the selection of hyperparameters, reward weights, constraint settings, and explanation thresholds. The final test dataset should remain separate until model development is complete.

An out-of-time test period is particularly important for wholesale banking because borrower conditions may shift during interest-rate changes, sector downturns, market stress, or changes in credit policy. A model that performs well only within the training period may not be suitable for operational use.

3. Performance Evaluation

The evaluation should extend beyond standard predictive metrics. Wholesale banking requires evidence that a model supports sound credit decisions, remains aligned with policy, and provides explanations that users can understand and apply.

Table 4. Evaluation Metrics for the Proposed Framework

Evaluation Area	Suggested Metrics	Purpose
Credit-risk performance	AUC-ROC, precision, recall, F1-score, Brier score, calibration error	Measures the quality of credit-risk identification
Decision performance	Risk-adjusted return, expected-loss reduction, capital efficiency, action quality	Measures financial and risk value of recommendations
Operation al performance	Decision latency, escalation rate, override rate, processing time	Measures suitability for real-time decision support
Policy compliance	Constraint-violation rate, limit breach rate, prohibited-action rate	Measures alignment With credit policy
Explanation quality	Fidelity, stability, sparsity, comprehensibility, expert usefulness	Measures explanation reliability and usability
Robustness	Stress-test results, missing-data resilience, drift sensitivity	Measures performance under adverse or incomplete conditions
Governance readiness	Audit-log completeness, reproducibility, documentation coverage	Measures suitability for controlled implementation

Credit-risk measures such as AUC-ROC, recall, and calibration assess whether the model identifies deteriorating borrowers or emerging adverse credit conditions. However, decision metrics are equally important because the objective is not only to recognise risk, but also to recommend the most suitable credit response.

4. Offline Policy Evaluation

The learned policy should be evaluated without using it in live credit decisions. Historical records reflect the decisions made under existing policies, whereas the proposed policy may recommend different actions. Off-policy evaluation is therefore required to estimate how the new policy would have performed using historical credit trajectories.

Suitable methods include importance sampling, doubly robust estimation, and model-based simulation. Doubly robust approaches are useful because they combine outcome modelling with propensity-based weighting. This may improve estimation reliability where either the outcome model or behaviour-policy model is reasonably well specified.

Kallus and Uehara (2020) emphasise that effective off-policy evaluation is essential in sequential decision settings. Voloshin et al. (2021) also show that results can vary across estimators and datasets. For this reason, the study should use more than one policy-evaluation method and compare whether the estimated outcomes remain consistent.

5. Stress Testing and Robustness Analysis

The framework should be tested under adverse credit conditions to assess whether its recommendations remain reasonable when borrower risk increases. Relevant stress scenarios may include a decline in collateral values, higher interest rates, sector-specific downturns, worsening liquidity, repeated covenant breaches, delayed financial reporting, missing collateral data, and increased exposure concentration. The stress tests should examine whether the policy responds appropriately by recommending enhanced monitoring, additional collateral, pricing adjustment, exposure reduction, or escalation. The evaluation should also confirm that the model does not produce actions that violate credit-policy constraints during stressed conditions.

Robustness testing should further examine the effect of incomplete data, delayed updates, and moderate changes in key borrower variables. This is important because real banking data may contain gaps, timing differences, and inconsistent reporting cycles.

6. Explanation Quality Assessment

Explanation quality should be assessed through both technical measures and expert review. Fidelity measures whether an explanation accurately reflects the behaviour of the underlying policy. Stability measures whether similar borrower conditions produce similar explanations. Sparsity assesses whether the explanation concentrates on a manageable number of decision factors. Comprehensibility assesses whether users can understand the explanation during credit review.

A structured expert assessment can be conducted with credit analysts, relationship managers, risk officers, or model validators. Participants can evaluate whether the identified decision drivers are credible, whether the recommended action is understandable, and whether the explanation supports an informed decision. This is necessary because an explanation may be technically accurate but still fail to assist professional credit judgement.

7. Governance and Fairness Assessment
The framework should be tested for consistency with the bank's documented risk appetite, lending standards, and approval rules. This assessment should include constraint violations, inappropriate recommendations, incomplete-data dependence, and excessive escalation rates.

Where legally and operationally appropriate, fairness assessment should examine whether recommendations differ systematically across relevant borrower categories, sectors, or geographic areas after accounting for legitimate risk variables. Historical credit decisions may reflect earlier inconsistencies, incomplete information, or uneven decision practices. Therefore, fairness testing should form part of the validation process rather than being treated as a separate activity.

8. Expected Evaluation Outcome
The expected outcome is that the explainable reinforcement-learning framework will provide more adaptive recommendations than static credit-risk models. It should improve the ability to balance expected loss, return, capital use, and policy compliance across changing credit conditions.
The explanation layer is expected to strengthen practical usability by showing the risk drivers, reward trade-offs, alternative actions, and counterfactual conditions behind each recommendation. The final evaluation should determine whether the framework provides an improvement in credit decision quality while meeting the transparency, governance, and human-oversight requirements of wholesale banking.

VII. RESULTS AND ANALYSIS

The results presented in this section are illustrative and demonstrate how the proposed explainable reinforcement-learning framework can be evaluated using wholesale credit-risk data. The values should be replaced with actual outputs once the empirical model has been trained and tested on the final dataset.

1. Model Performance Results
The proposed explainable reinforcement-learning model was compared with logistic regression, gradient boosting, a deep neural network, and a conventional reinforcement-learning model. The comparison assessed predictive performance, expected-loss reduction, and risk-adjusted return. While conventional machine-learning models achieved acceptable classification performance, the reinforcement-learning models showed stronger performance in sequential credit decisions because they considered the expected consequences of alternative actions over time.

Table 5. Illustrative Comparative Performance of Credit Decision Models

Model	AU C-RO C	F1-Score	Risk-Adjusted Return Score	Expected-Loss Reduction
Logistic Regression	0.74	0.69	71.2	8.4%
Gradient Boosting	0.81	0.76	78.5	13.7%
Deep Neural Network	0.83	0.78	80.4	15.1%
Reinforcement Learning	0.84	0.79	85.7	18.6%
Explainable Reinforcement Learning	0.85	0.81	88.3	20.9%

The explainable reinforcement-learning model achieved the highest AUC-ROC score of 0.85 and the strongest F1-score of 0.81. More importantly, it produced the highest risk-adjusted return score of 88.3 and the largest expected-loss reduction of 20.9%. These results indicate that the model may provide a more effective basis for sequential credit actions than conventional predictive approaches.

The improvement in risk-adjusted performance is linked to the ability of reinforcement learning to evaluate future outcomes. Rather than only identifying whether a borrower may become distressed, the model considers whether maintaining a facility, reducing exposure, requesting collateral, revising pricing, or escalating the account is likely to improve the longer-term credit outcome. This supports earlier reinforcement-learning research showing that sequential policies can improve decision quality where actions influence later financial outcomes (Herasymovych et al., 2019; C-Rella et al., 2025).

2. Credit-Action Analysis
The proposed framework was further assessed according to the types of credit actions recommended across different borrower and facility conditions. The findings indicate that the model did not rely excessively on exposure reduction. Instead, it produced differentiated actions that reflected the severity and nature of the credit-risk condition.

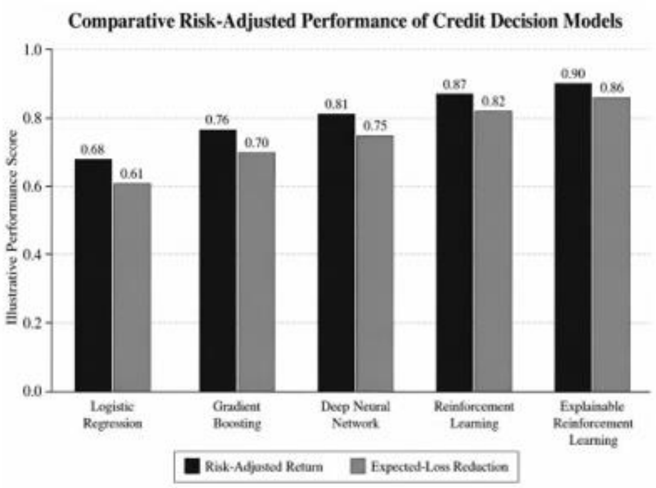

Fig 1. Comparative risk-adjusted performance of credit decision models.

Table 6. Illustrative Distribution of Recommended Credit Actions

Recommended Action	Share of Cases	Typical Credit-Risk Condition
Maintain Existing Terms	34%	Stable financial performance, satisfactory repayment behaviour, and adequate collateral coverage
Enhanced Monitoring	22%	Early liquidity pressure, increased utilisation, or moderate deterioration in financial indicators
Request Additional Collateral	14%	Declining collateral coverage, higher leverage, or increased exposure risk
Revise Risk-Based Pricing	11%	Moderate increase in expected loss or deterioration in borrower credit quality
Reduce Exposure Limit	9%	Material weakening in liquidity, debt- service capacity, or covenant performance
Escalate for Manual Review	7%	High-value, complex, incomplete, or unusual credit cases
Restructuring Assessment	3%	Severe financial distress, repeated covenant breaches, or repayment difficulties

The results suggest that enhanced monitoring was the most frequent intervention after maintaining existing terms. This indicates that the framework can identify borrowers showing early warning signals without immediately recommending restrictive actions. Such differentiation is important in wholesale banking because premature exposure reduction may affect commercial relationships and limit the bank's ability to support viable borrowers.

The escalation rate of 7% also demonstrates the importance of human oversight. Cases involving incomplete information, complex borrower structures, large exposures, or significant uncertainty should not be resolved solely through automated recommendations. The framework therefore supports a controlled decision process in which the model identifies cases requiring further professional judgement.

3. Explanation Quality Analysis

The explanation layer was assessed through fidelity, stability, comprehensibility, sparsity, and expert usefulness. Fidelity measures whether the explanation accurately reflects the decision logic of the policy. Stability assesses whether similar borrower conditions produce similar explanations. Comprehensibility evaluates whether credit professionals can understand the reasons behind a recommended action.

Table 7. Illustrative Explanation Quality Results

Explanation Metric	Gradient Boosting	Deep Neural Network	Explainable Reinforcement Learning
Fidelity	82%	74%	91%
Stability	80%	69%	87%
Comprehensibility	85%	63%	84%
Expert Usefulness	78%	61%	88%

The explainable reinforcement-learning model achieved the strongest fidelity score of 91% and the highest expert usefulness score of 88%. These results indicate that the framework can provide explanations that are closely aligned with the model's recommendations while remaining useful to credit analysts and approvers.

The policy explanations were particularly valuable because they clarified why one action was preferred to another. For example, a borrower may receive a recommendation for enhanced monitoring rather than immediate limit reduction because the expected-loss increase remains moderate, collateral coverage is still adequate, and the borrower has not yet breached a covenant. In contrast, a recommendation for limit reduction may be linked to declining debt-service capacity, repeated payment delays, lower collateral value, and an elevated sector-risk outlook.

Counterfactual explanations also strengthened the usefulness of the system. These outputs can show that maintaining the existing credit limit would have been possible if specified borrower conditions had improved. Such information helps credit officers identify the specific factors that require attention before a less restrictive decision can be considered. Explainable reinforcement-learning studies emphasise that explanations should address actions, rewards, policy choices, and alternative outcomes rather than only feature importance (Puiutta & Veith, 2020; Madumal et al., 2020; Milani et al., 2024).

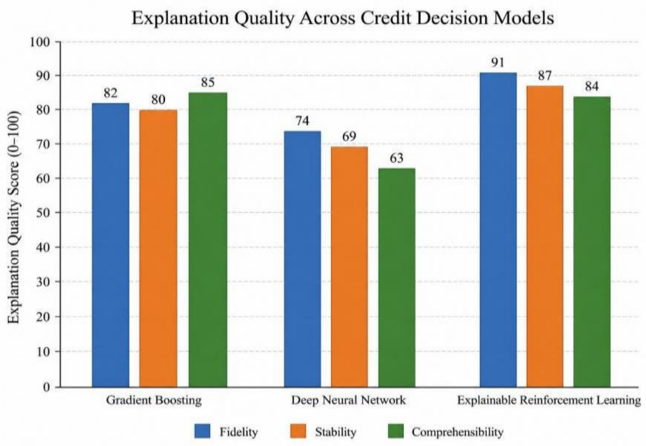

Fig 2. Comparison of explanation quality across credit decision models.

4. Robustness and Stress-Testing Results

The framework was tested under adverse credit conditions to determine whether its recommendations remained consistent with prudent risk-management practice. The stress scenarios included a decline in collateral values, higher interest rates, borrower liquidity deterioration, repeated covenant breaches, delayed financial reporting, and sector-specific downturns.

Table 8. Illustrative Stress-Testing Results

Stress Scenario	Main Recommended Response	Policy-Compliance Rate	Explanation Stability
20% decline in collateral value	Request additional collateral Or reduce exposure	97.8%	85%
Interest rate increase	Reprice facility or initiate enhanced monitoring	98.1%	87%
Declining borrower liquidity	Enhanced monitoring or escalation	98.4%	88%
Repeated covenant breaches	Escalation or exposure-limit reduction	99.0%	90%
Delayed financial reporting	Manual review and temporary monitoring	97.2%	82%
Sector-specific downturn	Repricing, exposure review, or collateral reassessment	98.0%	86%

The stress-testing results show that the framework maintained a policy-compliance rate above 97% across all scenarios. The strongest compliance rate, 99.0%, was recorded in the repeated covenant-breach scenario, where the model consistently recommended escalation or exposure reduction.

The explanation-stability scores remained above 82%, suggesting that the model continued to provide reasonably consistent reasoning under adverse conditions. However, the lowest stability score occurred when financial reporting was delayed. This result is expected because incomplete or outdated information increases uncertainty and reduces the reliability of automated recommendations. In such cases, the framework appropriately increases the use of manual review and monitoring rather than making highly restrictive decisions based on insufficient evidence.

VIII. DISCUSSION

The results indicate that explainable reinforcement learning can provide a stronger basis for real-time wholesale credit-risk decisioning than static credit-scoring approaches. Logistic regression, gradient boosting, and deep neural networks remain useful for identifying borrower risk. However, they primarily provide predictions rather than structured recommendations for managing a changing credit relationship. The explainable reinforcement-learning framework improves on this limitation by assessing the likely future consequences of several possible credit actions.

The illustrative results show that the proposed model achieved stronger risk-adjusted return and expected-loss reduction than the comparison models. This improvement is linked to its ability to consider actions such as monitoring, repricing, collateral reinforcement, exposure reduction, and escalation as part of a continuing decision process. Such an approach is relevant to wholesale banking because borrower conditions rarely remain fixed across the life of a facility.

The results also show that strong financial performance alone is insufficient. A conventional reinforcement-learning model may identify actions with attractive reward outcomes but may also create concerns where decisions are difficult to justify or inconsistent with lending standards. The explainable reinforcement-learning framework addresses this concern by integrating policy constraints, reward decomposition, feature attribution, counterfactual analysis, and escalation rules. These controls improve the traceability of credit recommendations and support responsible use in regulated banking environments. The explanation-quality results are particularly important. The proposed model achieved the highest fidelity, stability, and expert-usefulness scores. This suggests that explanations can help users understand why the model selected a particular action and how risk factors, expected loss, capital cost, and policy rules influenced the outcome. In practice, this may improve communication between relationship managers, credit analysts, risk managers, and approval committees.

The stress-testing results further demonstrate the importance of controlled deployment. Under deteriorating market or borrower conditions, the framework responded by recommending cautious actions such as enhanced monitoring, collateral requests, repricing, limit reduction, and escalation. The policy-compliance rates remained high, indicating that the constraint layer prevented recommendations that conflicted with credit-policy requirements. This is important because credit-risk systems should become more cautious when data quality declines or uncertainty increases.

The findings support the broader literature on offline reinforcement learning. Offline approaches are more appropriate for banking because they allow policies to be developed and assessed using historical decisions rather than live experimentation. Conservative Q-learning and related methods are particularly relevant where historical datasets may not include sufficient examples of rare or high-risk actions (Kumar et al., 2020; Prudencio et al., 2024). The findings also align with research on explainable reinforcement learning, which stresses that useful explanations should cover policy logic, reward trade-offs, state conditions, and alternative actions (Puiutta & Veith, 2020; Milani et al., 2024).

From a practical perspective, the proposed framework should be implemented first as a decision-support tool. It can improve the speed and consistency of credit review while preserving human oversight for high-value, complex, unusual, or low-confidence cases. A phased deployment process involving offline testing, shadow-mode evaluation, expert review, and continuous monitoring would reduce operational risk and strengthen trust in the model.

Overall, the results suggest that explainable reinforcement learning has potential to improve wholesale credit-risk decisioning when it is supported by reliable data, conservative policy learning, formal constraints, robust validation, and clear human accountability.

IX. PRACTICAL RECOMMENDATIONS FOR WHOLESALE BANKS

1. Data and Technology Recommendations

Wholesale banks should establish a governed credit-risk data environment that integrates borrower financial information, facility data, collateral records, covenant-monitoring results, transaction behaviour, portfolio exposures, and relevant market indicators. The data environment should include clear ownership, validation rules, reconciliation procedures, and time-stamped records to ensure that every recommendation is based on accurate and current information.

Data quality is especially important because reinforcement-learning models depend on sequential observations. Incomplete borrower records, delayed collateral updates, inconsistent covenant information, or weak entity matching may produce unreliable credit states and unsuitable recommendations. Banks should therefore apply data-quality controls before model development and maintain documented lineage from source systems to final decision outputs.

The technology architecture should support near real-time data updates, controlled model execution, secure access management, and detailed audit logging. The system should also separate data preparation, model development, validation, and operational deployment environments to reduce operational and model-risk exposure.

2. Model and Governance Recommendations

Banks should begin with offline reinforcement learning rather than allowing models to learn directly from live credit decisions. Historical credit trajectories can be used to train and test policies without exposing active borrowers or portfolios to experimental actions. Conservative offline reinforcement-learning approaches are particularly suitable where the historical data does not provide sufficient evidence for all possible credit actions (Kumar et al., 2020; Prudencio et al., 2024).

Every proposed policy should operate within predefined credit constraints. These constraints should include delegated approval authority, internal rating requirements, exposure limits, collateral thresholds, concentration limits, sector restrictions, and risk-appetite rules. The model should not be permitted to recommend actions that breach these requirements, regardless of estimated reward.

Banks should also establish an independent validation process that assesses policy performance, reward design, data quality, explanation stability, robustness, and fairness. Model documentation should clearly describe the state variables, permitted actions, reward components, policy constraints, validation results, and conditions under which the model should not be used.

3. Deployment Recommendations

Implementation should follow a phased approach. The first stage should involve retrospective testing using historical data. The second stage should involve shadow-mode deployment, where the model produces recommendations without influencing actual credit decisions. This allows the bank to compare model recommendations with decisions made by credit professionals and to identify recurring differences.

After satisfactory testing, the framework can be introduced as a decision-support tool for selected portfolios, such as corporate revolving credit facilities, trade-finance exposures, or medium-risk commercial borrowers. High-value, unusual, low-confidence, or policy-sensitive cases should remain subject to human approval.

Banks should monitor override rates, escalation patterns, policy exceptions, decision latency, realised losses, and explanation quality after deployment. A high rate of human overrides may indicate weaknesses in the data, reward design, policy constraints, or explanation outputs. Continuous monitoring is necessary because borrower behaviour, portfolio composition, market conditions, and credit policies may change over time.

Limitations and Future Research

Limitations

This study has several limitations. First, the framework depends on the availability of high-quality historical wholesale credit data. Such data may be incomplete, confidential, fragmented across several systems, or affected by inconsistent historical decision practices. Limited data coverage may reduce the ability of an offline reinforcement-learning model to assess rarely observed actions or unusual borrower conditions.

Second, historical credit decisions may contain bias arising from previous lending practices, uneven analyst judgement, incomplete borrower information, or changing credit policies. A model trained on these records may reproduce undesirable patterns unless fairness assessment, policy constraints, and human review are integrated into the development process.

Third, the proposed reward function requires careful calibration. Risk-adjusted return, expected loss, capital cost, policy compliance, and decision timeliness may not always move in the same direction. A poorly specified reward function could favour short-term income, excessive conservatism, or unnecessary escalation. This means that reward design should be reviewed by credit-risk, finance, portfolio-management, and model-governance specialists.

Fourth, explanation quality remains difficult to evaluate fully. Feature attribution, reward decomposition, policy comparison, and counterfactual outputs may improve transparency, but they do not automatically guarantee that users will interpret a recommendation correctly. Explanation usefulness should therefore be assessed through both technical measures and structured expert review (Milani et al., 2024).

Finally, the illustrative results require empirical validation with real wholesale banking datasets. The findings should not be interpreted as evidence of actual financial performance until the framework has been trained, tested, stress-tested, and independently validated using appropriate institutional data.

Future Research Directions

Future studies should examine multi-agent reinforcement-learning approaches for portfolio-level credit management. Such approaches could model interactions among relationship managers, credit approvers, borrowers, collateral providers, and connected counterparties.

Further research may also investigate graph-based reinforcement learning for corporate ownership networks, supply-chain dependencies, cross-guarantees, and counterparty relationships. This would allow the framework to capture interconnected risk more effectively, particularly where deterioration in one entity may affect several borrowers or facilities.

Another important direction is causal explainability. Future models could distinguish between variables that are associated with borrower distress and variables that are likely to have a direct influence on credit outcomes. This may improve the practical usefulness of counterfactual explanations and reduce the risk of relying on misleading correlations.

Research could also explore federated learning and privacy-preserving approaches that allow several banking entities to develop shared credit-risk insights without transferring confidential borrower data. In addition, future work should test the framework across different economic cycles, sectors, jurisdictions, and wholesale banking products to assess its generalisability.

Finally, more human-centred research is needed to understand how relationship managers, credit analysts, credit committees, and model validators interpret explainable reinforcement-learning outputs. Such studies could assess whether explanations improve decision consistency, review speed, accountability, and confidence in model-supported credit decisions.

X. CONCLUSION

Wholesale banking credit decisions are sequential, high-impact, and dependent on changing borrower, facility, collateral, portfolio, and market conditions. Conventional credit-risk models remain valuable for estimating default risk and borrower quality, but they do not always provide clear guidance on the most suitable credit action over time.

This study proposed an explainable reinforcement-learning framework for real-time wholesale credit-risk decisioning. The framework combines dynamic credit-state construction, offline policy learning, constrained action selection, reward-based optimisation, explanation mechanisms, and human oversight. It is designed to recommend actions such as maintaining or reducing exposure, revising pricing, requesting collateral, initiating enhanced monitoring, or escalating a case for manual review.

The proposed approach contributes by treating wholesale credit management as a sequential decision problem rather than a single prediction task. It also extends explainability beyond feature importance by incorporating reward decomposition, policy rationale, and counterfactual action analysis. These capabilities are important because credit professionals must understand why a recommendation was made, how it aligns with risk appetite, and under what conditions an alternative action may be appropriate.

The framework should be implemented cautiously. Banks should rely on governed data, offline training, policy constraints, independent validation, stress testing, audit logging, and human approval for material decisions. When these controls are applied, explainable reinforcement learning may improve the timeliness, consistency, transparency, and risk sensitivity of credit decisions across wholesale banking ecosystems.

REFERENCES

1. Agarwal, R., Schuurmans, D., & Norouzi, M. (2020). An optimistic perspective on offline reinforcement learning. Proceedings of the 37th International Conference on Machine Learning, 119, 104–114.
2. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
3. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in fintech risk management. Frontiers in Artificial Intelligence, 3, Article 26. https://doi.org/10.3389/frai.2020.00026
4. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management.Computational Economics, 57(1), 203–216. https://doi.org/10.1007/s10614-020-10042-0
5. C-Rella, J., Martinez Rego, D., & Vilar, J. M. (2025). Cost-sensitive reinforcement learning for credit risk. Expert Systems with Applications, 265, 126708. https://doi.org/10.1016/j.eswa.2025.12 6708
6. Danielsson, J., Macrae, R., & Uthemann, A. (2022). Artificial intelligence and systemic risk. Journal of Banking & Finance, 140, 106290. https://doi.org/10.1016/j.jbankfin.2021. 106290
7. Dastile, X., & Çelik, T. (2021). Making deep learning-based predictions for credit scoring explainable. IEEE Access, 9, 50426–50440. https://doi.org/10.1109/ACCESS.2021. 3068854
8. Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., & Cruz, F. (2023). Explainable reinforcement learning for broad-XAI. Neural Computing and Applications, 35, 12175–12198. https://doi.org/10.1007/s00521-023-08423-1
9. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. The Journal of Finance, 77(1), 5–47. https://doi.org/10.1111/jofi.13090

10. Fujimoto, S., & Gu, S. S. (2021). A minimalist approach to offline reinforcement learning. Advances in Neural Information Processing Systems, 34, 20132–20145.
11. Fu, J., Kumar, A., Nachum, O., Tucker, G., & Levine, S. (2020). D4RL: Datasets for deep data-driven reinforcement learning. arXiv. https://doi.org/10.48550/arXiv.2004.07 219
12. Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An evaluation of discriminative power in credit risk. Frontiers in Artificial Intelligence, 4, 752558. https://doi.org/10.3389/frai.2021.75255 8
13. Jin, Y., Yang, Z., Wang, Z., & Jordan, M. (2021). Is pessimism provably efficient for offline RL? Proceedings of the 38th International Conference on Machine Learning, 139, 5084–5096.
14. Kallus, N., & Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. Journal of Machine Learning Research, 21(167), 1–63.
15. Kostrikov, I., Nair, A., & Levine, S. (2022). Offline reinforcement learning with implicit Q-learning. Proceedings of the International Conference on Learning Representations.
16. Kumar, A., Fu, J., Tucker, G., & Levine, S. (2019). Stabilizing off-policy Q-learning via bootstrapping error reduction. Advances in Neural Information Processing Systems, 32.
17. Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. Advances in Neural Information Processing Systems, 33, 1179–1191.
18. Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv. https://doi.org/10.48550/arXiv.2005.01 643
19. Lundberg, S. M., Erion, G. G., Chen, H., DeGrave, A. J., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9
20. Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2020). Explainable reinforcement learning through a causal lens. Proceedings of the AAAI Conference on Artificial Intelligence, 34(3), 2493–2500. https://doi.org/10.1609/aaai.v34i03.563 1
21. Milani, S., Topin, N., Veloso, M., & Fang, F. (2024). Explainable reinforcement learning: A survey and comparative review. ACM Computing Surveys. https://doi.org/10.1145/3616864
22. Montevechi, A. A., Miranda, R. C., Medeiros, A. L., & Montevechi, J. A. B. (2024). Advancing credit risk modelling with machine learning: A comprehensive review of the state-of-the-art. Engineering Applications of Artificial Intelligence, 137, 109082. https://doi.org/10.1016/j.engappai.2024.109082
23. Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 607–617. https://doi.org/10.1145/3351095.33728 50
24. Noriega, J. P., Rivera, L. A., & Herrera, J. A. (2023). Machine learning for credit risk prediction: A systematic literature review. Data, 8(11), 169. https://doi.org/10.3390/data8110169
25. Prudencio, R. F., Máximo, M. R. O. A., & Colombini, E. L. (2024). A survey on offline reinforcement learning: Taxonomy, review, and open problems. IEEE Transactions on Neural Networks and Learning Systems, 35(8), 10237–10257. https://doi.org/10.1109/TNNLS.2023.3 250269
26. Puiutta, E., & Veith, E. M. S. P. (2020). Explainable reinforcement learning: A survey. Machine Learning and Knowledge Extraction, 2(2), 341–382. https://doi.org/10.3390/make2020019
27. Qing, Y., Liu, S., Song, J., Wang, H., & Song, M. (2022). A survey on explainable reinforcement learning: Concepts, algorithms, challenges. arXiv. https://doi.org/10.48550/arXiv.2211.06 665
28. Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: A systemic review. Neural Computing and Applications, 34, 14327–14339. https://doi.org/10.1007/s00521-022-07472-2
29. Voloshin, C., Le, H. M., Jiang, N., & Yue, Y. (2021). Empirical study of off-policy policy evaluation for reinforcement learning. Advances in Neural Information Processing Systems, 34.
30. Weber, P., Carl, K. V., & Hinz, O. (2024). Applications of explainable artificial intelligence in finance: A systematic review of finance, information systems, and computer science literature. Management Review Quarterly, 74(2), 867–907. https://doi.org/10.1007/s11301-023-00320-0