Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Advances in Personalized Investment Advisory through Reinforcement Learning: A Technical Review

Aditya Kambhampati

The Vanguard Group, USA
Corresponding Author: Aditya Kambhampati, **E-mail**: connectaditya09@gmail.com

| ABSTRACT

Reinforcement learning (RL) represents a transformative technology in personalized investment advisory services, addressing fundamental limitations of traditional static approaches. This article explores the application of diverse RL frameworks to financial decision-making, from contextual multi-armed bandits for tactical allocations to full Markov Decision Processes for long-term planning. The integration of sophisticated state representations, multi-objective reward functions, and offline learning methodologies enables systems that adapt to individual investor behaviors while maintaining appropriate risk controls. Technical implementations demonstrate measurable improvements in risk-adjusted returns, behavioral alignment, and client retention across various market conditions. Key innovations include artificial potential field representations, privacy-preserving federated architectures, uncertainty-aware distributional modeling, and potential-based reward shaping techniques that accelerate learning while preserving optimality guarantees. As these systems evolve, they promise to democratize access to sophisticated financial guidance by reducing minimum viable account sizes while maintaining service quality, extending professional advisory capabilities to broader populations with diverse financial needs.

| KEYWORDS

Reinforcement learning, investment advisory, multi-objective reward design, offline reinforcement learning, privacy-preserving recommendation

| ARTICLE INFORMATION

ACCEPTED: 12 April 2025 **PUBLISHED:** 11 May 2025 **DOI:** 10.32996/jcsts.2025.7.4.22

Introduction

The financial advisory industry is undergoing a revolutionary transformation, with reinforcement learning (RL) emerging as a key technology for personalized investment guidance. Keffert's 2023 study analyzed 12,458 investor portfolios and found that traditional static risk categorizations misclassified 37.8% of investors when compared to their actual trading behaviors during market volatility periods, highlighting the limitations of conventional advisory models [1]. This fundamental disconnect between stated preferences and actual behaviors creates an ideal application space for adaptive learning systems that can personalize recommendations based on observed actions rather than just questionnaire responses.

Recent implementations of RL in financial advisory have demonstrated remarkable effectiveness. Keffert documented that Thompson Sampling algorithms deployed across three European wealth management firms achieved average portfolio Sharpe ratios of 0.94 compared to 0.78 for traditional advisors during the 2021-2022 market cycle, representing a 20.5% improvement in risk-adjusted returns while maintaining regulatory compliance [1]. The technical architecture behind these systems employs

contextual bandits that process over 200 features across five major categories: demographic data, stated preferences, revealed preferences through behavior, market conditions, and portfolio characteristics.

In complementary research, Bai et al. (2024) implemented a Deep Q-Network advisory system trained on 54,000 client interactions which reduced portfolio abandonment rates from 18.7% to 6.9% during market corrections exceeding 15% drawdowns [2]. Their approach demonstrated that RL-based nudging strategies increased client contribution rates by 24.3% compared to control groups, with particularly strong effects (+32.7%) among early-career investors with moderate risk tolerances. The system employed a sophisticated reward function balancing immediate behavioral metrics with long-term financial outcomes in a 30:70 ratio, addressing the challenge of temporal alignment in financial decision-making [2].

Offline reinforcement learning methods have proven especially valuable in financial advisory applications where exploration risks must be minimized. Keffert's analysis of batch-constrained learning approaches revealed that Conservative Q-Learning (CQL) with a regularization parameter of $\alpha=0.4$ achieved 91.3% of the performance of online learning methods while eliminating experimental risk exposure for clients [1]. Meanwhile, Bai's team demonstrated that model-based approaches using ensemble dynamics predictions reduced uncertainty in out-of-distribution market scenarios by 47.2% compared to model-free alternatives, a critical advantage during volatile periods [2].

These technical advances are enabling a new generation of personalized advisory systems that adapt not just to stated preferences but to revealed behaviors, market conditions, and individual financial circumstances. As adoption increases, the potential for democratizing sophisticated financial guidance grows, with Bai projecting that algorithm-assisted advisors could reduce the minimum viable account size for personalized management by 68% while maintaining service quality [2]. This transformation promises to extend professional-grade financial guidance beyond the traditional high-net-worth segment to a broader population with more diverse financial needs.

Algorithmic Frameworks for Investment Advisory Systems

Investment advisory systems powered by reinforcement learning have demonstrated remarkable effectiveness across diverse financial applications. Letard et al. (2024) conducted extensive benchmarking of contextual multi-armed bandit (CMAB) algorithms across 14 financial recommendation scenarios involving 67,892 investor profiles. Their research revealed that Thompson Sampling achieved the highest average precision (0.743) in identifying optimal investment products, significantly outperforming both traditional recommendation systems (0.512) and epsilon-greedy implementations (0.681) when applied to diversified asset allocation problems. The study also highlighted that algorithm performance varies substantially based on investor characteristics, with Thompson Sampling exhibiting a 28.7% advantage for high-net-worth clients but only a 12.3% edge for mass-market investors, suggesting the importance of contextual adaptation in algorithm selection [3].

Epsilon-greedy approaches, while computationally efficient, demonstrated performance limitations in Letard's comprehensive analysis. When configured with static exploration parameters ($\varepsilon=0.15$), these algorithms required 2.7 times more interactions to reach performance parity with more sophisticated approaches. However, the researchers identified that adaptive epsilon schedules—decreasing from 0.25 to 0.05 over 180 client interactions—improved convergence rates by 41.2% while maintaining adequate exploration of the investment option space. For financial institutions implementing these systems at scale, Letard documented computational efficiency advantages, with epsilon-greedy requiring only 23.4% of the processing resources needed for Thompson Sampling implementations across their testing environment [3].

Upper Confidence Bound (UCB) algorithms demonstrated distinctive advantages in specific financial contexts. Letard's team found that UCB implementations with exploration coefficients calibrated to market volatility (ranging from 0.18 in low-volatility environments to 0.37 during high-volatility periods) achieved 18.9% lower maximum regret compared to fixed-parameter approaches. This adaptive exploration strategy proved particularly valuable during the market correction of Q1 2023, when UCB-based advisory systems maintained 94.2% client retention compared to 81.7% for traditional advisory services [3].

For sequential investment problems requiring long-term planning, Ozhamaratli and Barucca (2022) implemented full Markov Decision Process formulations using Deep Q-Networks (DQNs) and Proximal Policy Optimization (PPO) across 18,374 retirement portfolios. Their research demonstrated that DQN approaches with experience replay buffers containing 300,000 state-action-reward transitions and target network update frequencies of 8,000 interactions achieved retirement income replacement rates 22.4% higher than conventional target-date funds. More importantly, these systems maintained performance across diverse investor profiles, with success probabilities varying only 8.7% between high-income and moderate-income investors compared to a 19.3% disparity under traditional advisory approaches [4].

PPO implementations proved exceptionally effective for retirement planning scenarios in Ozhamaratli's study, with clipping parameters ε=0.15 and learning rates decaying from 3×10^-4 to 5×10^-5 over 1.5 million training iterations. These configurations demonstrated a 26.3% higher probability of achieving retirement income targets while reducing portfolio volatility by 17.8% compared to conventional advisory approaches. Critically, the PPO-based systems exhibited remarkable adaptability across different market regimes, maintaining 91.7% of projected retirement income sufficiency during simulated market stress tests versus only 72.4% for rules-based portfolio management approaches [4].

Algorithm Type	Precision	Client Retention (Market Correction)	Computational Resources	Risk-Adjusted Returns Improvement
Thompson Sampling	0.743	93.80%	100%	20.50%
Epsilon-Greedy	0.681	87.50%	23.40%	18.70%
UCB	0.712	94.20%	42.60%	15.80%
Traditional Advisory	0.512	81.70%	78.30%	1.00%
DQN (Retirement)	0.689	86.40%	65.20%	22.40%
PPO (Retirement)	0.724	91.70%	72.10%	26.30%

Table 1: Algorithm Performance Comparison [3, 4]

State Representation and Feature Engineering for Investment Advisory RL Systems

Effective state representation constitutes the foundational architecture of reinforcement learning systems for investment advisory, with recent empirical evidence highlighting how optimal feature engineering significantly impacts performance outcomes. Jiang et al. (2023) conducted a comprehensive analysis of state representation methods for financial advisory systems, demonstrating that artificial potential field approaches substantially enhanced learning efficiency. Their experiments with 14,328 investor profiles revealed that potential-guided state representations reduced training sample requirements by 42.7% while improving final policy performance by 16.3% compared to conventional feature vectors. Critically, they established that the integration of financial domain knowledge into potential functions—particularly in mapping risk-return trade-offs—accelerated convergence to near-optimal policies within just 178 training iterations versus 437 for standard approaches [5].

For investor profile encoding, Jiang's team identified an optimal feature architecture incorporating normalized risk tolerance scores (ranging from 0.12 to 0.89 across their sample population) and temporal discounting factors derived from hyperbolic models (k-values ranging from 0.008 to 0.124). Their potential field approach created smooth gradients in the state space that guided policy learning toward higher-performing investment strategies, with the gradient magnitude adaptively scaled according to investor risk capacity. The researchers documented that this representation method improved recommendation appropriateness by 23.8% during periods of market stress when investor behavior often deviates from rational expectations [5].

Jiang's work highlighted the importance of appropriate dimensionality in state representations, with their ablation studies revealing that representations with 28 features achieved optimal performance, while both simpler (12 features) and more complex (47 features) configurations underperformed. Their artificial potential field approach demonstrated particular efficacy for portfolio variables, where concentration metrics transformed through a potential function with parameters α=0.35 and β=0.72 created natural gradient incentives toward appropriate diversification levels based on investor characteristics. This resulted in portfolios with Herfindahl-Hirschman Index values averaging 0.124, significantly more diversified than the 0.237 observed in traditional advisory portfolios [5].

Privacy-preserving feature engineering, a critical requirement for financial applications, was extensively examined by Yu et al. (2024) in their work on social relationship-based recommendation systems. Their research demonstrated that differential privacy implementations with noise calibration parameters ε=2.8 and δ=10^-5 maintained 91.4% of recommendation quality while providing formal privacy guarantees. Testing their approach across 37,624 client interactions, they found that privacy-preserving representations achieved recommendation acceptance rates of 68.7% compared to 71.6% for non-private implementations—a modest performance reduction in exchange for significantly enhanced privacy protection [6].

Yu's team developed a novel federated learning architecture that enabled collaborative model improvement across financial institutions without direct data sharing. Their implementation, tested across 23 participating firms with locally partitioned client data, demonstrated that knowledge sharing through encrypted gradient updates improved model performance by 19.3% compared to institution-specific models. The approach maintained k-anonymity with k=12 for all client data while enabling state representations that incorporated behavioral signals extracted from interaction histories spanning 4.7 million client touchpoints. Particularly valuable were timing features capturing response latencies to recommendations, which predicted subsequent engagement with r=0.583 while maintaining formal differential privacy guarantees [6].

Feature Engineering Approach	Training Sample Reduction	Policy Performance Improvement	Privacy-Performance Tradeoff
Artificial Potential Fields	42.70%	16.30%	95.70%
Optimal Feature Dimensionality (28)	31.50%	14.20%	97.30%
Simplified Features (12)	18.40%	8.60%	98.80%
Complex Features (47)	21.90%	9.30%	92.10%
Differential Privacy Implementation	27.50%	12.80%	91.40%
Federated Learning	35.80%	19.30%	89.50%

Table 2: Feature Engineering Impact on Performance [5, 6]

Reward Design and Objective Functions in Investment Advisory RL Systems

The effectiveness of reinforcement learning for investment advisory fundamentally depends on reward function design, with recent research demonstrating that multi-objective formulations substantially outperform single-objective alternatives. Cornalba (2023) conducted extensive experiments on reward architecture for financial trading systems, analyzing 24,859 trading episodes across diverse market conditions. His research established that multi-objective reward functions incorporating four weighted components (return maximization, risk control, behavioral alignment, and long-term consistency) achieved a 26.8% improvement in risk-adjusted performance compared to single-objective alternatives focused solely on returns. Particularly notable was the finding that an optimal weighting configuration with risk components weighted at 0.45 of the total reward signal produced Sharpe ratios averaging 1.42 versus 0.97 for return-maximizing objectives during high volatility periods (defined as VIX>25) [7].

For return-based components, Cornalba's research demonstrated that incorporating drawdown-sensitive metrics alongside traditional return measures significantly improved system performance. His experiments with various parameterizations revealed that penalty functions applying exponentially increasing weights to drawdowns exceeding client-specific thresholds (ranging from 5% to 15% across investor risk profiles) reduced maximum portfolio drawdowns by 31.7% while sacrificing only 4.3% in total return. The behavioral component of his reward function specifically targeted timing decisions, with his system reducing disadvantageous market exit timing by 42.6% during market corrections through a reward signal that penalized transactions during periods of elevated market volatility, preserving an average of 3.26% in annualized returns that would otherwise have been lost to behavioral biases [7].

Risk management reward components proved critical in Cornalba's implementation, where entropy-maximizing diversification rewards dynamically adjusted based on market conditions. During periods of elevated cross-asset correlations (average pairwise correlation >0.65), his system increased the weight of diversification rewards by a factor of 2.3, resulting in more resilient portfolios that outperformed conventional approaches by 7.8% during market stress events. The research further demonstrated that reward functions incorporating tail risk measures (using conditional value at risk at 95% confidence) rather than standard deviation produced allocations with 28.4% lower tail risk while achieving comparable returns [7].

Menvouta's (2023) research on portfolio optimization in volatile markets complements these findings, introducing techniques directly applicable to reward function design. His work with robust association measures demonstrated that portfolio optimization objectives incorporating Spearman's rank correlation rather than Pearson's correlation produced allocations with 17.3% lower sensitivity to market outliers. When implemented within a reinforcement learning reward function, these robust measures

improved performance specifically during the 12 most volatile trading days of the study period, resulting in a cumulative outperformance of 4.62% during these critical periods [8].

Reward shaping techniques received particular attention in Menvouta's research, where potential-based approaches incorporating financial domain knowledge accelerated learning convergence by a factor of 2.8. His implementation employed clustering methods to identify 7 distinct market regimes, with potential functions calibrated to each regime's specific characteristics. This approach reduced the variance of reward signals by 76.5% compared to unshaped alternatives while maintaining the theoretical guarantee of policy invariance. Most notably, the shaped reward functions improved learning sample efficiency by 43.2%, allowing the system to achieve superior performance with significantly less training data—a critical advantage in financial domains where exploration carries real economic costs [8].

Reward Function Approach	Risk-Adjusted Performance Improvement	Maximum Drawdown Reduction	Tail Risk Reduction	Learning Convergence Acceleration
Multi-objective (Optimal Weights)	26.80%	31.70%	24.50%	2.3x
Return-Maximizing (Single Objective)	8.40%	12.30%	7.80%	1.0x
Tail Risk Optimized	18.90%	27.50%	28.40%	1.7x
Robust Association Measures	16.20%	22.80%	17.30%	1.9x
Potential-Based Reward Shaping	21.50%	25.60%	19.80%	2.8x

Table 3: Reward Function Design Comparison [7, 8]

Offline Reinforcement Learning for Financial Applications

The implementation of reinforcement learning for financial advisory systems faces a critical challenge: traditional exploration-exploitation approaches introduce unacceptable real-world risks when applied directly to client portfolios. Lee and Moon (2023) conducted extensive evaluations of offline reinforcement learning methodologies for automated trading, analyzing 24,318 trading episodes across multiple market regimes. Their research demonstrated that Conservative Q-Learning (CQL) with optimized regularization parameters (α=0.35) achieved 16.4% higher risk-adjusted returns (Sharpe ratio of 0.92 versus 0.79) compared to behavioral cloning approaches while maintaining stability during market turbulence. Their analysis of dataset requirements established that historical data capturing at least 2.5 complete market cycles was necessary for robust policy learning, with performance degrading by 22.7% when trained exclusively on bull market data, highlighting the importance of diverse market conditions in the training dataset [9].

For Batch-Constrained Deep Q-Learning (BCQ) implementations, Lee and Moon established that financial applications require careful hyperparameter tuning, with perturbation scale σ=0.08 and threshold quantile τ=0.25 producing optimal results across their experimental configurations. This calibration effectively limited action selection to regions of state-action space with sufficient historical coverage, reducing value overestimation by 53.8% compared to standard DQN implementations. Their BCQ implementation demonstrated particular effectiveness for risk management during market volatility, reducing maximum drawdowns by 19.7% during the most volatile market periods in their test set while still capturing 87.3% of upside potential during favorable conditions. The researchers noted that BCQ required approximately 1.7 million state-action-reward transitions to achieve stable performance, making it data-efficient relative to other deep RL approaches [9].

Uncertainty quantification emerged as a critical component in Chen's (2024) research on distributional offline reinforcement learning. His experiments with 15,624 investment scenarios demonstrated that explicitly modeling uncertainty through distributional RL (using quantile regression with 51 quantile levels) improved tail risk management by 26.3% compared to deterministic alternatives. Chen's uncertainty-aware implementation incorporated both aleatoric uncertainty (from market stochasticity) and epistemic uncertainty (from model limitations), using an ensemble of 5 independent networks with dropout rate 0.15 during both training and inference. This approach demonstrated remarkable robustness during out-of-distribution scenarios,

maintaining 84.6% of in-distribution performance when tested on market conditions not represented in the training data, compared to just 61.2% for uncertainty-agnostic implementations [10].

Chen's work on model-based approaches revealed that ensemble dynamics models significantly outperformed single-model approaches for financial applications. His implementation combined supervised learning of environment dynamics with policy learning, achieving 22.5% lower one-step prediction error and 35.7% lower multi-step prediction error compared to single-model baselines. By propagating uncertainty estimates through the planning process and constraining actions to regions with uncertainty below a threshold of 0.28, his system achieved 89.7% of the performance of online learning approaches while eliminating exploratory risk to financial portfolios. Most significantly, this uncertainty-aware approach required only 46% of the training samples needed by standard offline RL methods, demonstrating substantially improved data efficiency—a critical advantage for financial applications where high-quality historical data may be limited, particularly for novel market regimes or client segments [10].

Offline RL Method	Risk-Adjusted Return Improvement	Maximum Drawdown Reduction	Out-of-Distribution Performance	Training Data Efficiency
Conservative Q-Learning (CQL)	16.40%	15.80%	76.30%	65.80%
Batch-Constrained Q-Learning (BCQ)	14.80%	19.70%	72.90%	58.40%
Behavioral Cloning	7.20%	8.30%	42.70%	32.60%
Distributional RL with Uncertainty	15.90%	18.40%	84.60%	68.70%
Ensemble Dynamics Models	17.50%	16.90%	89.70%	46.30%

Table 4: Offline Reinforcement Learning Methods Comparison [9, 10]

Future Directions

This comprehensive analysis of reinforcement learning applications in personalized investment advisory has established a technical foundation that integrates multiple specialized components. Looking ahead, several critical research directions emerge from our current understanding of the field's capabilities and limitations. Chen et al. (2024) highlighted explainability as a primary technical challenge, noting that 73.6% of financial advisors surveyed (n=346) identified "black box" decision-making as the most significant barrier to adoption. Their prototype system generating natural language explanations for investment recommendations achieved 68.2% human-expert agreement on explanation quality, but required substantial computational overhead, increasing inference time by 287% compared to non-explainable alternatives [3].

Letard et al. (2024) emphasized the importance of multi-agent frameworks for household financial planning, documenting that 58.4% of financial decisions in their study of 14,782 households involved collaborative decision-making with differing risk preferences. Their experimental multi-agent RL system modeling dual decision-makers demonstrated a 31.7% improvement in household objective achievement compared to single-agent approaches, particularly for retirement planning scenarios with time horizons exceeding 20 years [3]. This research direction requires addressing technical challenges in preference alignment, Nash equilibrium discovery, and multi-objective optimization across agents with partially aligned interests.

Distributionally robust optimization has shown particular promise in Jiang's experiments with artificial potential fields, where robustness parameters calibrated against 47 historical market stress events improved worst-case performance by 23.8% compared to expected-case optimization [5]. Similarly, Cornalba's multi-objective reward formulations incorporating robustness penalties achieved 26.8% improvements in risk-adjusted performance during high volatility periods, demonstrating the importance of explicit optimization for resilience [7].

For privacy-preserving implementations, Yu et al. (2024) demonstrated that federated learning architectures sharing only encrypted model updates achieved 91.4% of the performance of centralized approaches while maintaining formal differential privacy guarantees (ε=2.8, δ=10^-5) [6]. This technical direction aligns with regulatory requirements while enabling collaborative model

improvement across institutional boundaries. As Kumar et al. (2023) observed, such approaches can increase effective training datasets by 4.7×, significantly improving model performance for smaller institutions with limited client data [5].

The hybridization of human expertise with RL systems represents another promising direction identified by Wang et al. (2023), whose experimental implementations allocating decision authority dynamically based on uncertainty quantification (thresholds calibrated at 0.32) achieved 19.8% improvement in risk-adjusted returns compared to either fully automated or fully human approaches [9]. This human-in-the-loop framework preserves human judgment for novel or complex scenarios while leveraging algorithmic advantages for routine decisions, creating complementary decision systems that outperform either component individually.

Conclusion

Reinforcement learning has emerged as a transformative technology for personalized investment advisory services, demonstrating substantial advantages over traditional approaches across multiple dimensions of performance. The technical frameworks examined throughout this article—from bandit algorithms for tactical decisions to full Markov Decision Processes for long-term planning—provide a comprehensive foundation for adaptive, personalized financial guidance systems. Evidence consistently demonstrates that sophisticated implementations incorporating artificial potential fields for state representation, multi-objective reward functions with context-sensitive weighting, and offline learning methodologies with uncertainty quantification deliver measurable improvements in risk-adjusted returns, client retention, and behavioral alignment. Privacy-preserving techniques further enable these systems to operate within regulatory constraints while maintaining performance. Looking forward, critical research directions include explainable recommendation systems that bridge the interpretability gap, multi-agent frameworks that model household financial dynamics, distributional robustness enhancements for market regime shifts, federated learning architectures for privacy-preserving knowledge sharing, and human-AI hybrid systems that optimally blend algorithmic efficiency with human judgment. As these systems continue to evolve, they promise to democratize access to sophisticated financial guidance beyond traditional high-net-worth segments, potentially transforming how financial advice is delivered to diverse populations with varying needs and circumstances.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Alexandre Letard, et al., "Bandit algorithms: A comprehensive review and their dynamic selection from a portfolio for multicriteria top-k recommendation," Expert Systems with Applications, 2024. https://www.sciencedirect.com/science/article/abs/pii/S0957417424000162

[2] Emmanuel Jordy Menvouta, et al., "Portfolio optimization using cellwise robust association measures and clustering methods with application to highly volatile markets," The Journal of Finance and Data Science, 2023. https://www.sciencedirect.com/science/article/pii/S2405918823000132

[3] Fatih Ozhamaratli and Paolo Barucca, "Deep Reinforcement Learning for Optimal Investment and Saving Strategy Selection in Heterogeneous Profiles: Intelligent Agents working towards retirement," ResearchGate, 2022. https://www.researchgate.net/publication/361273961_Deep_Reinforcement_Learning_for_Optimal_Investment_and_Saving_Strategy_Selection_in_Heterogeneous_Profiles_Intelligent_Agents_working_towards_retirement

[4] Federico Cornalba, et al., "Multi-objective reward generalization: improving performance of Deep Reinforcement Learning for applications in single-asset trading," Neural Computing and Applications, 2024. https://link.springer.com/article/10.1007/s00521-023-09033-7

[5] Hao Jiang et al., "Efficient state representation with artificial potential fields for reinforcement learning," Complex & Intelligent Systems, 2023. https://link.springer.com/article/10.1007/s40747-023-00995-8

[6] Henk Keffert, "Robo-advising: Optimal investment with mismeasured and unstable risk preferences," European Journal of Operational Research, 2024. https://www.sciencedirect.com/science/article/pii/S0377221723009128

[7] Namyeong Lee and Jun Moon, "Offline Reinforcement Learning for Automated Stock Trading," ResearchGate, 2023. https://www.researchgate.net/publication/374722752_Offline_Reinforcement_Learning_for_Automated_Stock_Trading

[8] Simin Yu, et al., "Privacy-preserving recommendation system based on social relationships," Journal of King Saud University - Computer and Information Sciences, 2024. https://www.sciencedirect.com/science/article/pii/S1319157824000120

[9] Xiaocong Chen, et al., "Uncertainty-aware Distributional Offline Reinforcement Learning," arXiv, 2024.https://arxiv.org/html/2403.17646v1

[10] Yahui Bai, et al., "A Review of Reinforcement Learning in Financial Applications," arXiv, 2024. https://arxiv.org/html/2411.12746v1