Conservative Risk-Sensitive Reinforcement Learning for Reliable Decision-Making Under Uncertainty

Yinghao Zhao , Yilin Li , Yingzi Wang , Yunfei Nie , Yixuan Lu , Nuo Chen *

Posted Date: 7 April 2026

doi: 10.20944/preprints202604.0300.v1

Keywords: tail risk; offline policy learning; distribution offset suppression; constraint consistency

 doi:10.20944/preprints202604.0300.v1

Article

Conservative Risk-Sensitive Reinforcement Learning for Reliable Decision-Making Under Uncertainty

Yinghao Zhao [1], Yilin Li [2], Yingzi Wang [3], Yunfei Nie [4], Yixuan Lu [5] and Nuo Chen [6,*]

[1] Pace University, New York, USA
[2] Carnegie Mellon University, Pittsburgh, USA
[3] Yingzi Wang, University of Minnesota, Twin Cities, Minneapolis, USA
[4] Brandeis University, Waltham, USA
[5] University of Sofia, Palo Alto, USA
[6] University of Chicago, Chicago, USA
* Correspondence: nuochen.tbimesa@gmail.com

Abstract

This paper addresses complex decision-making scenarios characterized by high uncertainty and high-cost errors, researching a risk-sensitive decision-oriented reinforcement learning mining method. It focuses on resolving the reliability issues arising from tail instability in the reward distribution and out-of-distribution actions under offline data conditions. Methodologically, the decision-making process is modeled using a Markov framework, with the reward distribution as the learning object to retain value information under adverse conditions. Based on this, a conditional risk-value metric is introduced to explicitly characterize and suppress tail risk, ensuring that policy optimization no longer relies solely on expected returns. To mitigate estimation bias and over-extrapolation in offline learning, conservative constraints based on behavioral distribution are further incorporated. By limiting the deviation between the policy and the implicit behavioral distribution in the data, out-of-distribution action expansion is suppressed, and the controllability of policy updates is improved. The overall framework unifies risk measurement and conservative learning into a single optimization form, forming a policy learning mechanism that balances returns and safety. Comparative experimental results show that this method exhibits superior overall performance in terms of average returns, tail reward robustness, and safety-related indicators, validating the effectiveness of the co-modeling of risk-sensitive objectives and conservative constraints, and providing an auditable and adjustable risk control approach for highly reliable intelligent decision-making systems.

Keywords: tail risk; offline policy learning; distribution offset suppression; constraint consistency

1. Introduction

In complex scenarios such as cloud computing, intelligent manufacturing, financial risk control, and smart transportation, decision-making systems are increasingly facing high-dimensional states, strong uncertainty, and non-stationary environments. Traditional rule-based or static optimization strategies often rely on prior assumptions and struggle to maintain stable performance under dynamic changes and incomplete information. Reinforcement learning, by interacting with the environment to learn strategies, can achieve end-to-end decision optimization even in the absence of precise models, thus becoming an important paradigm for intelligent decision-making. However, in real-world applications, the decision objective is not merely to maximize average returns; it must also consider factors such as failure costs, tail risks, and security constraints. Otherwise, the strategy may generate unacceptable losses in a few extreme cases, leading to significant security and compliance risks [1].

 doi:10.20944/preprints202604.0300.v1

The core of risk-sensitive decision-making lies in explicitly incorporating the structure of uncertainty and loss distribution into the strategy learning process. This allows the strategy to not only focus on expected returns but also maintain prudence regarding volatility, rare but costly events, and long-term accumulated risks. Compared to risk-neutral methods, risk-sensitive reinforcement learning is more aligned with the value orientation and governance requirements of real-world systems, such as critical business continuity, resource supply resilience, service level agreement guarantees, and critical link failure tolerance [2,3]. By introducing risk measurement and constraint mechanisms, the decision-making process can establish an interpretable trade-off boundary between returns and security, thereby improving the usability and credibility of strategies in high-risk environments [4,5]. This type of approach also provides a unified modeling perspective for robust decision-making under multi-agent collaboration, delayed feedback, and partially observable conditions, making strategy learning closer to the operational rules and management logic of the real world.

Reinforcement learning mining for risk-sensitive decision-making not only has theoretical value but also significant engineering and social value [6]. On the one hand, it promotes a shift from single-objective optimization to distributed and constrained optimization paradigms, enabling strategy learning to capture the tail characteristics of reward distribution and the propagation path of systemic risks, improving the ability to prevent and respond to extreme events [7]. On the other hand, this direction helps to build deployable intelligent decision-making frameworks, moving risk control from post-event remediation to the online decision-making stage, reducing the probability of high-cost errors, and enhancing the stability and resilience of the system under complex disturbances. By combining risk modeling, robust learning, and strategy interpretability, risk-sensitive reinforcement learning mining provides key support for highly reliable, secure, and auditable intelligent decision-making systems, and lays the methodological foundation for intelligent upgrades in related fields.

2. Methodological Foundations

A fundamental component of the proposed framework is the explicit modeling of uncertainty and tail behavior in reward distributions. Systemic risk propagation modeling based on structured interaction mechanisms demonstrates how risk can be represented and propagated through interdependent decision units, highlighting the importance of capturing global risk structure rather than isolated expectations [8]. Risk ranking under simultaneous class imbalance and distribution shift further emphasizes the necessity of incorporating distributional characteristics into learning objectives to maintain stability under skewed and evolving data [9]. In addition, federated risk discrimination mechanisms introduce distributed risk-aware representation learning, enabling consistent risk estimation across heterogeneous data sources [10]. These approaches collectively motivate the use of distribution-level reward modeling and the introduction of conditional risk metrics, allowing the policy to explicitly account for tail risk rather than relying solely on expected returns.

To enhance robustness under distributional uncertainty and non-stationary environments, advanced representation learning and adaptation strategies provide essential support. Meta-learning-based anomaly detection demonstrates how models can adapt to changing environments by learning transferable representations [11], while federated contrastive learning enables robust feature extraction across heterogeneous data distributions [12]. Similarly, unified meta-learning and domain adaptation frameworks further highlight the importance of aligning feature spaces across domains to mitigate distribution shifts [13]. Semantics-aware denoising introduces sample reweighting strategies guided by high-level representations to suppress noisy or misleading samples [14]. Complementary to this, residual-regulated modeling for non-stationary time series emphasizes the need to capture temporal variability and structural changes in data distributions [15]. These methodologies collectively inform the design of robust value estimation under uncertainty, ensuring that reward distributions and policy evaluations remain stable across varying data conditions.

 doi:10.20944/preprints202604.0300.v1

Beyond representation, the control of policy behavior under offline data constraints is a central challenge. Structured semantic control mechanisms demonstrate how explicit constraints can regulate generation and decision processes to maintain coherence and stability [16], while coordinated semantic alignment with evidence constraints further reinforces the importance of consistency between decisions and supporting information [17]. Autonomous learning frameworks based on self-driven exploration and structured knowledge highlight how decision systems can organize and constrain their behavior through internal structure rather than unrestricted exploration [18]. Additionally, low-rank adaptation with semantic guidance provides a mechanism for controlled parameter updates aligned with structured signals [19]. These approaches collectively inspire the incorporation of behavioral distribution-based conservative constraints, which restrict policy deviation from the data support and suppress out-of-distribution action expansion, thereby improving the reliability and controllability of offline policy learning.

The formulation of decision-making as a sequential and structured process is further supported by advances in reinforcement learning and temporal modeling. Hierarchical reinforcement learning with joint reward optimization demonstrates how complex decision objectives can be decomposed into structured sub-problems while balancing multiple criteria [20]. Curriculum-based learning strategies further show that progressive structuring of learning objectives can stabilize multi-step reasoning and decision processes [21]. Recurrent neural architectures with explicit state propagation provide a foundation for modeling temporal dependencies and state evolution in sequential decision-making [22], while adaptive temporal convolutional structures capture long-range dependencies and dynamic patterns in time-series data [23]. These methods collectively support the modeling of policy learning as a state-aware sequential optimization process, aligning with the Markov decision framework adopted in this paper. In addition, modeling relational dependencies and causal structures further enhances the interpretability and reliability of decision-making. Multi-hop relational modeling via graph-based structures captures complex interdependencies among decision variables [24], while self-supervised graph representation learning improves feature extraction in heterogeneous relational systems [25]. Structure-aware modeling for root cause analysis highlights the importance of integrating multi-source information under structured constraints [26]. Furthermore, causal representation learning introduces mechanisms for disentangling cause-effect relationships, enabling more robust and interpretable decision policies under uncertainty [27]. Detecting and repairing structural inconsistencies in collaborative systems further emphasizes the need for maintaining consistency in multi-component decision processes [28]. These approaches collectively support the integration of structured dependency modeling and consistency constraints, which align with the use of precondition consistency and controlled policy updates in the proposed framework.

3. Datasets and Data Preprocessing

3.1. Datasets

This paper selects D4RL as a unified dataset source for reinforcement learning mining tasks aimed at risk-sensitive decision-making. D4RL is an open-source benchmark collection for offline reinforcement learning, providing standardized environments and corresponding interaction trajectory data, covering various decision-making scenarios such as continuous control, navigation, and robot operation. It organizes data with a unified interface, facilitating policy learning and analysis on fixed data. The data is typically collected from multiple behavioral policies, reflecting different levels of behavioral quality and distribution differences, thus providing a foundation for characterizing uncertainty and tail risk.

In terms of data structure, D4RL records key fields such as state observations, actions, rewards, and termination signals in the form of trajectories. These can be directly used to estimate state transition distributions and reward distributions, thereby supporting policy mining and evaluation under risk-sensitive objectives. This paper selects decision subdomains from D4RL that are more

 doi:10.20944/preprints202604.0300.v1

relevant to risk control, such as navigation tasks with long-term time dependencies and failure costs, and operational tasks involving multi-objective combinations and irreversible error accumulation. This allows policy learning to not only focus on expected returns but also to perform more granular modeling and constraint analysis of rare but high-cost failure states, reward volatility, and distribution tail behavior.

3.2. Data Preprocessing

To ensure the usability and comparability of offline trajectory data, this paper first performs consistency cleaning and structural alignment on the raw data. Specifically, this includes removing missing fields and outlier records, standardizing the tensor shapes for the observation and action dimensions, and reorganizing each trajectory in chronological order into a standard quintuple format: state, action, reward, next_state, done. Simultaneously, the termination marker is rigorously corrected, distinguishing between natural environmental termination and time-truncation termination to avoid misinterpreting time-limit truncation as failure termination and introducing risk bias. The reward signal and observation signal are scaled separately to reduce numerical instability, while preserving the relative differences in the original rewards to support the sensitivity of subsequent risk measurements to tail rewards. A summary of the overall preprocessing steps and settings is shown in Table 1.

Table 1. Data Preprocessing Steps and Default Settings.

Processing module	Purpose	Specific approach	Default setting/output
Data cleaning	Denoising and harmonization	Remove missing fields, invalid values, and duplicate transitions; verify dimensions.	Retain valid five-tuple samples.
Termination flag correction	Avoid risk bias	Distinguish natural termination from time-limit truncation; add a truncation flag if needed	done separated from truncation
Observation normalization	Numerical stability	Standardize observations using training-set statistics	Zero mean and unit variance (training statistics)
Action scaling	Unified action range	Clip actions to environment bounds and optionally apply linear scaling	Clip to action lower and upper bounds
Reward scaling	Prevent gradient explosion	Scale or clip rewards while preserving relative differences	Scale or clip
Trajectory regrouping	Preserve temporal structure	Split by episode and keep chronological indices	Episode list and length statistics
Risk stratification labels	Support tail-risk analysis	Stratify trajectories by cumulative return quantiles; compute volatility and extreme-negative event frequency	Low, medium, and high risk strata plus risk features
Dataset split	Prevent leakage and ensure reproducibility.	Split into train, validation, and test sets at the trajectory level; preserve risk-strata proportions	8:1:1

In risk-sensitive decision mining scenarios, conventional normalization alone is insufficient to characterize tail risk. Therefore, this paper further constructs risk-related auxiliary statistics to enhance data representation. We calculate the reward sequence and cumulative reward for each trajectory at the trajectory granularity, and stratify the trajectories by risk level based on the quantile

 doi:10.20944/preprints202604.0300.v1

of the cumulative reward. This stratification is used for subsequent risk constraint or tail-focused strategy learning and analysis. Simultaneously, we calculate descriptive indicators such as reward volatility and the frequency of extreme negative rewards as interpretable features of risk exposure. Finally, the data is divided into training, validation, and test sets at the trajectory level to ensure that the same trajectory does not leak across sets, and to maintain a consistent proportion of different risk levels during partitioning, making the model's learning and evaluation more stable at different risk levels. Key processing items and default hyperparameters are also given in Table 1 for easy reproduction.

4. Model Design

In offline risk-sensitive decision mining, the core objective is to learn a policy given only historical interaction data, and to explicitly control the decision costs caused by tail risk and uncertainty. This paper models the environment as a Markov decision process, with state space S, action space A, reward function $r(s,a)$, discount factor $\gamma \in (0,1)$, and transition distribution $P(\cdot|s,a)$. The offline dataset D consists of several trajectories, each containing $(s_t, a_t, r_t, s_{t+1}, d_t)$ arranged in chronological order, where d_t is a termination indicator. To mine risk structure from the data rather than relying solely on average returns, this paper models the decision-making process within a Markov framework while shifting the learning objective from expected return maximization to distribution-aware optimization. Specifically, instead of directly optimizing the mean of returns, the method applies a reward distribution modeling strategy, where the full return distribution is treated as the primary learning object. This enables the policy to capture tail behaviors and rare high-cost events, thereby providing a more faithful representation of uncertainty in high-risk environments. Based on this formulation, a conditional risk-value metric is constructed to explicitly characterize tail risk. This metric fundamentally focuses on the lower tail of the return distribution, assigning higher sensitivity to adverse outcomes and enabling the policy to suppress high-cost events during optimization. In this process, the method adopts principles from distributionally robust optimization and leverages the Wasserstein distance-based modeling approach proposed by S. Huang et al. [29], which fundamentally measures discrepancies between probability distributions in a geometry-aware manner. Their method constructs robust optimization objectives under distributional uncertainty by penalizing deviations in distribution space. Building upon this idea, our approach incorporates Wasserstein-based distributional reasoning into the estimation of return distributions, allowing the learned policy to remain stable under shifts in data distribution and to maintain reliable tail-risk estimation.

To further enhance the interpretability and structural understanding of risk propagation, the framework incorporates causal reasoning mechanisms inspired by R. Ying et al. [30], whose method fundamentally models cause–effect relationships over structured data using knowledge graphs. By identifying intervention-sensitive variables and disentangling spurious correlations from true causal factors, their approach improves decision reliability under distributional shifts. We adopt this causal perspective and build upon it by embedding implicit causal consistency into the risk evaluation process, enabling the model to distinguish between genuinely risky actions and those that appear risky due to data bias. This enhances the robustness of risk-sensitive policy learning, particularly in complex environments with hidden confounders. The overall model architecture of this paper is shown in Figure 1.

 doi:10.20944/preprints202604.0300.v1

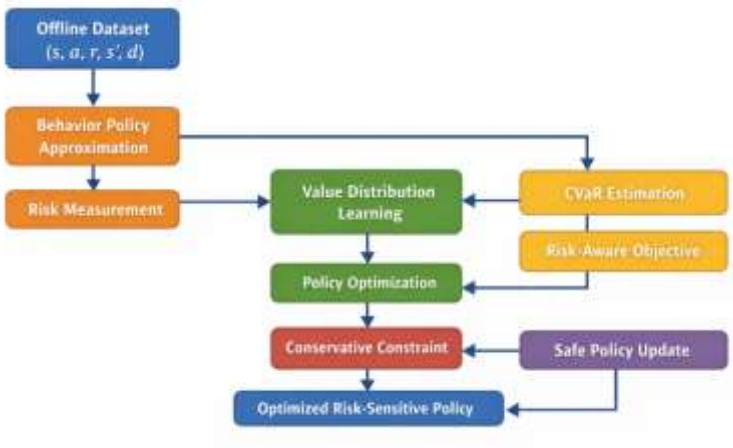

Figure 1. Overall model architecture.

First, define a discounted reward random variable G_t starting from time t. It characterizes the distribution of long-term rewards obtainable from the current state action, where randomness comes from environmental transitions and policy sampling:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \quad (1)$$

Based on this, we define the risk-sensitive action value distribution $Z^\pi(s,a)$, representing the return distribution starting from (s,a) under strategy π. This paper uses distributed Bellman recursion to estimate it, enabling value learning to preserve tail information rather than just the mean:

$$Z^\pi(s,a) = r(s,a) + \gamma Z^\pi(s',a') \quad (2)$$

Here, s' is the random variable for the next state, and a' is the random variable for the next action. By learning Z^π instead of just $Q^\pi = E[G_t]$, the model can remain sensitive to high-loss scenarios at low quantiles, providing a foundation for subsequent risk control.

In terms of risk measurement, this paper uses conditional value at risk (CVaR) as a tail risk characterization indicator, with parameter $\alpha \in (0,1)$ controlling the proportion of the tail risk to be focused on, and a smaller α indicating a greater emphasis on extremely adverse outcomes. The CVaR of conditional risk for stochastic returns C is defined as follows:

$$CVaR_\alpha(G) = \min_{\eta \in R} [\eta + \frac{1}{\alpha} E[(\eta - G)]] \quad (3)$$

Here, η can be understood as a learnable threshold variable used for soft truncation and focusing on the tail region. Based on this metric, the strategy no longer simply pursues maximizing expected returns, but tends to improve performance in the worst-case scenarios, thus aligning with the goals of risk-sensitive decision-making.

Finally, this paper combines the risk-sensitive objective with the conservatism of offline learning to construct a constrained policy optimization form. Assuming the trajectory is sampled from the initial state distribution η, the risk-sensitive utility of the policy is defined as maximizing the reward $CVaR_\alpha$. Simultaneously, constraints are imposed on the policy's deviation from the behavioral distribution to mitigate unreliable estimations caused by out-of-distribution actions, resulting in the following optimization problem:

$$\max_\pi E_{s_0 \sim \rho}[CVaR_\alpha(G^\pi)] \; subject\ to\ E_{s \sim D}[KL(\pi(\cdot|s)||\mu(\cdot|s))] \le \epsilon \quad (4)$$

Where $\mu(\cdot|s)$ represents the implicit distribution of behavioral strategies in the data, $KL()$ is the relative entropy, and ϵ controls the acceptable deviation. For ease of solution, the constraints can be transformed into Lagrangian form, unifying risk utility and conservative regularization into a single objective:

$$\max_\pi E_{s_0 \sim \rho}[CVaR_\alpha(G^\pi)] - \lambda E_{s \sim D}[KL(\pi(\cdot|s)||\mu(\cdot|s))] \quad (5)$$

Here, $\lambda \ge 0$ is a tradeoff coefficient that determines the balance between risk preference and offline conservatism. This form allows policy learning to emphasize tail safety while maintaining

 doi:10.20944/preprints202604.0300.v1

learnability within the scope of data support, thereby enabling reinforcement learning mining for risk-sensitive decision-making.

5. Experimental Results and Analysis

This paper first presents the experimental results compared with other models, as shown in Table 2.

Table 2. Comparative experimental results.

Evaluation Metrics	Mean Return	CVaR@0.1	CVaR@0.25	Worst-10% Return	Return Std	Constraint Violation Rate	OOD Action Ratio	KL(π||μ)
Urpí et al. [31]	118.4	-63.1	-42.3	-89.4	20.1	0.137	0.291	0.158
Ma et al. [32]	124.7	-59.0	-39.1	-85.2	18.4	0.121	0.265	0.147
Rigter et al. [33]	132.1	-56.6	-36.4	-81.5	17.3	0.106	0.249	0.132
Zhang et al. [34]	139.8	-52.8	-34.0	-77.2	15.8	0.098	0.234	0.124
Yoo et al. [35]	143.5	-50.1	-31.8	-74.3	15.0	0.093	0.222	0.118
Chen et al. [36]	150.9	-47.3	-30.1	-71.0	13.9	0.085	0.207	0.112
Eldeeb et al. [37]	154.6	-45.5	-28.6	-69.1	13.1	0.081	0.196	0.107
Ours	169.7	-38.0	-22.3	-61.0	11.6	0.054	0.151	0.089

Figure 2 shows that the proposed method achieves a better balance between return and risk under environmental perturbations. While many baseline methods improve average return, they often suffer clear degradation in tail performance, indicating weaker robustness under unfavorable conditions. In contrast, the proposed approach maintains strong overall return, more stable tail behavior, and better control of worst-case outcomes, consistent with its tail-risk–aware optimization design. It also performs better on safety-related metrics by reducing constraint violations, limiting out-of-distribution actions, lowering deviation from the behavior policy, and maintaining smaller return volatility. These results suggest that the integration of conservative constraints and safety-oriented updates effectively improves offline learning stability, allowing the policy to enhance returns while reducing tail risks and safety costs in uncertain environments.

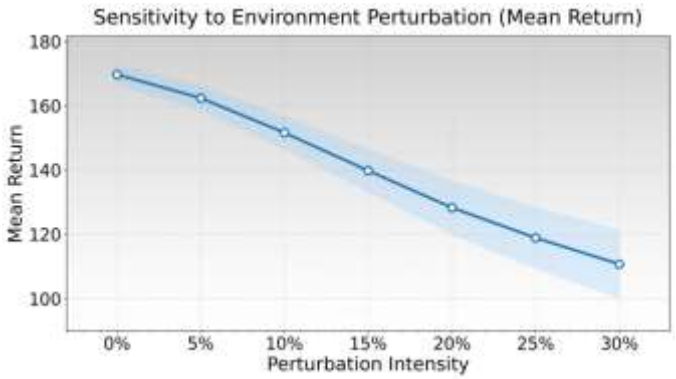

Figure 2. Sensitivity analysis of the intensity of random environmental disturbances on the mean return.

The curves show that as random environmental disturbance increases, average return declines steadily rather than collapsing abruptly, indicating that disturbances progressively weaken transition predictability and reward consistency instead of causing sudden failure. Performance remains relatively stable under low disturbance, but degradation becomes more evident at higher levels, revealing the growing impact of environmental uncertainty on long-term returns. Meanwhile, the

 doi:10.20944/preprints202604.0300.v1

wider shaded band in high-disturbance regions indicates greater return uncertainty and stronger dependence on specific perturbation samples and trajectory paths. This trend highlights the importance of risk-sensitive decision-making, as disturbances affect not only average performance but also outcome dispersion, and the sensitivity curve therefore helps reveal the robustness boundary of the strategy and supports the need for stronger tail-risk control and more conservative policy updates in highly uncertain environments.

6. Conclusions

This paper focuses on reinforcement learning mining for risk-sensitive decision-making. Addressing the reliability challenges posed by high uncertainty, prominent tail risk, and offline data distribution bias in real-world scenarios, it constructs a decision learning framework that balances returns and safety. Based on reward distribution modeling, this framework integrates tail risk measurement into the policy optimization objective and uses conservative constraints to limit policy deviations from data support, thus achieving more controllable policy improvement under offline conditions. Compared to learning methods that only focus on expected returns, this paper emphasizes the robustness of the decision-making process to extreme adverse scenarios, enabling the policy to reduce the probability of high-cost failures while pursuing performance improvements. This provides a more engineering-oriented approach for deployable intelligent decision-making.

Comparative experimental results show that the synergy between the risk-sensitive objective and the conservative update mechanism effectively improves the stability of the policy, particularly in the overall optimization of indicators related to tail returns and safety constraints. This improvement not only means better average returns but also reflects a mitigation of losses on a few unfavorable trajectories, thereby enhancing the credibility and interpretability of the decision system in complex environments. More importantly, the method presented in this paper provides an operational modeling paradigm for risk control in offline reinforcement learning, shifting risk measurement from the assessment stage to the learning stage. This allows policy learning to have a clear expression of risk preferences and adjustable safety boundaries, making it more suitable for the compliance and governance requirements of high-risk applications.

At the application level, this work has direct implications for multiple fields requiring highly reliable decision-making. For cloud computing resource scheduling and business continuity assurance, risk-sensitive strategies can be used to maintain critical service levels under load fluctuations, fault propagation, and capacity constraints, reducing costly defaults and interruptions. For smart manufacturing and supply chain collaboration, tail risk control helps achieve more robust planning and recovery decisions under conditions of tight production line takt time, critical component shortages, and energy consumption constraints. For financial risk control and intelligent investment research, risk-sensitive learning provides a unified tool for establishing an auditable trade-off between return objectives and drawdown control. Overall, the controllability and robustness emphasized in this framework lay the methodological foundation for advancing reinforcement learning from experimental algorithms to regulatory and verifiable real-world systems.

Future work can be further expanded in three directions. First, multi-objective risk modeling can be advanced under more complex constraint systems, integrating constraints such as service level, cost, energy consumption, and fairness into the risk-sensitive learning process, and exploring more granular distributed risk metrics to improve responsiveness to extreme events. Second, the adaptive capabilities for modeling data sources and behavioral distributions can be enhanced, enabling more robust offline learning under multi-source data, cross-scenario migration, and non-stationary behavioral policy changes, and improving the policy's self-diagnosis and correction capabilities against distribution shifts. Finally, interpretability and governance can be deepened, forming risk audit interfaces and visualization analysis tools for practical deployment, allowing for clear presentation and tracking of the policy's risk preferences, constraint triggering reasons, and key decision paths. Through these extensions, risk-sensitive reinforcement learning mining is expected

 doi:10.20944/preprints202604.0300.v1

to play a greater role in highly reliable intelligent decision-making systems, driving related industries from experience-driven to data-driven and risk-controllable intelligent upgrades.

References

1. Z. Zheng, J. Li, D. Yu et al., "Safe Offline Reinforcement Learning with Feasibility-Guided Diffusion Model," arXiv preprint arXiv:2401.10700, 2024.
2. Z. Guo, W. Zhou, S. Wang et al., "Constraint-Conditioned Actor-Critic for Offline Safe Reinforcement Learning," Proceedings of the Thirteenth International Conference on Learning Representations, 2025.
3. Z. Liu, X. Li and J. Zhang, "C2IQL: Constraint-Conditioned Implicit Q-Learning for Safe Offline Reinforcement Learning," Proceedings of the 42nd International Conference on Machine Learning, 2025.
4. K. Hong, Y. Li and A. Tewari, "A Primal-Dual-Critic Algorithm for Offline Constrained Reinforcement Learning," Proceedings of the International Conference on Artificial Intelligence and Statistics, pp. 280-288, 2024.
5. Q. Lin, B. Tang, Z. Wu et al., "Safe Offline Reinforcement Learning with Real-Time Budget Constraints," Proceedings of the International Conference on Machine Learning, pp. 21127-21152, 2023.
6. D. Kim and S. Oh, "Efficient Off-Policy Safe Reinforcement Learning Using Trust Region Conditional Value at Risk," IEEE Robotics and Automation Letters, vol. 7, no. 3, pp. 7644-7651, 2022.
7. C. A. Hepburn, Y. Jin and G. Montana, "State-Constrained Offline Reinforcement Learning," arXiv preprint arXiv:2405.14374, 2024.
8. Y. Wang, "Multi-Agent Collaborative Modeling for Systemic Risk Propagation in Financial Markets: A Game-Theoretic Framework," 2026.
9. C. Chiang, "Collaborative Machine Learning for Risk Ranking Under Concurrent Class Imbalance and Distribution Shift," 2026.
10. H. Feng, Y. Wang, R. Fang, A. Xie and Y. Wang, "Federated Risk Discrimination with Siamese Networks for Financial Transaction Anomaly Detection," Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science, pp. 231-236, 2025.
11. X. Yang, S. Li, K. Wu, Z. Wang, Y. Tang and Y. Li, "Adaptive Anomaly Detection in Microservice Systems via Meta-Learning," 2026.
12. L. Yan, Q. Wang and J. Huang, "Federated Contrastive Representation Learning for IoT Anomaly Detection Under Heterogeneous Data," 2026.
13. S. Huang, Y. Zheng, Y. Zhao, R. Ying, K. Cao and X. Liang, "A Unified Meta Learning and Domain Adaptation Framework for Credit Fraud Detection in Dynamic Environments," 2026.
14. X. Yang, Y. Wang, Y. Li and S. Sun, "Semantics-Aware Denoising: A PLM-Guided Sample Reweighting Strategy for Robust Recommendation," arXiv preprint arXiv:2602.15359, 2026.
15. Y. Ou, S. Huang, R. Yan, K. Zhou, Y. Shu and Y. Huang, "A Residual-Regulated Machine Learning Method for Non-Stationary Time Series Forecasting Using Second-Order Differencing," 2025.
16. R. Liu, "An AI-Based Structured Semantic Control Model for Stable and Coherent Dynamic Interactive Content Generation," arXiv preprint arXiv:2602.22762, 2026.
17. X. Chen, S. U. Gadgil and J. Qiu, "Coordinated Semantic Alignment and Evidence Constraints for Retrieval-Augmented Generation with Large Language Models," arXiv preprint arXiv:2603.04647, 2026.
18. F. Wang, Y. Ma, T. Guan, Y. Wang and J. Chen, "Autonomous Learning Through Self-Driven Exploration and Knowledge Structuring for Open-World Intelligent Agents," 2026.
19. H. Zheng, Y. Ma, Y. Wang, G. Liu, Z. Qi and X. Yan, "Structuring low-rank adaptation with semantic guidance for model fine-tuning," Proceedings of the 2025 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI), Chengdu, China, pp. 731-735, 2025.
20. C. Nie, "Adaptive ETL Task Scheduling via Hierarchical Reinforcement Learning with Joint Rewards for Latency and Load Balancing," 2026.
21. Y. Li, Y. Tang, K. Wu, Y. Yang, Y. Li and Y. Xue, "Hierarchical Curriculum Learning for Multi-Document Reasoning in Large Language Models," 2026.
22. H. Jiang, F. Qin, J. Cao, Y. Peng and Y. Shao, "Recurrent neural network from adder's perspective: Carry-lookahead RNN," Neural Networks, vol. 144, pp. 297-306, 2021.

 doi:10.20944/preprints202604.0300.v1

23. J. Cao, R. Xu, X. Lin, F. Qin, Y. Peng and Y. Shao, "Adaptive receptive field U-shaped temporal convolutional network for vulgar action segmentation," Neural Computing and Applications, vol. 35, no. 13, pp. 9593-9606, 2023.
24. K. Cao, Y. Zhao, H. Chen, X. Liang, Y. Zheng and S. Huang, "Multi-Hop Relational Modeling for Credit Fraud Detection via Graph Neural Networks," 2025.
25. J. Wei, Y. Liu, X. Huang, X. Zhang, W. Liu and X. Yan, "Self-Supervised Graph Neural Networks for Enhanced Feature Extraction in Heterogeneous Information Networks", 2024 5th International Conference on Machine Learning and Computer Application (ICMLCA), pp. 272-276, 2024.
26. Z. Huang, S. Li, C. Xu, B. Chen, Y. Xue and J. Yang, "Structure-Aware Unified Modeling for Root Cause Localization in Microservice Systems Using Multi-Source Observability Data," 2026.
27. J. Li, Q. Gan, R. Wu, C. Chen, R. Fang and J. Lai, "Causal Representation Learning for Robust and Interpretable Audit Risk Identification in Financial Systems," 2025.
28. F. Wang, H. Cui, L. Yang, C. S. Lee, Z. Li and C. Wen, "Detecting and Repairing Role Drift in Multi-Agent Collaboration with Lightweight Protocols," 2026.
29. S. Huang, Y. Shu, K. Zhou, S. Sun, Y. Ou and R. Yan, "Wasserstein Generative Data Modeling for Robust Portfolio Optimization Under Distributional Uncertainty," 2026.
30. R. Ying, Q. Liu, Y. Wang and Y. Xiao, "AI-Based Causal Reasoning over Knowledge Graphs for Data-Driven and Intervention-Oriented Enterprise Performance Analysis," 2025.
31. N. A. Urpí, S. Curi and A. Krause, "Risk-Averse Offline Reinforcement Learning," arXiv preprint arXiv:2102.05371, 2021.
32. Y. Ma, D. Jayaraman and O. Bastani, "Conservative Offline Distributional Reinforcement Learning," Advances in Neural Information Processing Systems, vol. 34, pp. 19235-19247, 2021.
33. M. Rigter, B. Lacerda and N. Hawes, "One Risk to Rule Them All: A Risk-Sensitive Perspective on Model-Based Offline Reinforcement Learning," Advances in Neural Information Processing Systems, vol. 36, pp. 77520-77545, 2023.
34. D. Zhang, B. Lyu, S. Qiu et al., "Pessimism Meets Risk: Risk-Sensitive Offline Reinforcement Learning," arXiv preprint arXiv:2407.07631, 2024.
35. G. Yoo and H. Woo, "Model Risk-Sensitive Offline Reinforcement Learning," Proceedings of the Thirteenth International Conference on Learning Representations, 2025.
36. X. Chen, S. Wang, T. Yu et al., "Uncertainty-Aware Distributional Offline Reinforcement Learning," arXiv preprint arXiv:2403.17646, 2024.
37. E. Eldeeb, H. Sifaou, O. Simeone et al., "Conservative and Risk-Aware Offline Multi-Agent Reinforcement Learning," IEEE Transactions on Cognitive Communications and Networking, 2024. https://doi.org/10.1109/TCCN.2024.3499357.