Reinforcement of Learning Theoretical Foundations and Exploration of Multi-Domain Applications

Yitao Lin

Petroleum Institute, China University of Petroleum (Beijing) At Karamay, Karamay, Xinjiang Uygur Autonomous Region, China

Abstract. Reinforcement learning, as an important branch of artificial intelligence, with its unique mechanism of learning optimal strategies through the interaction of intelligence with the environment, makes it show great application potential and significance in many fields. This paper focuses on exploring the applications of reinforcement learning in five fields, including game AI, robotics, autonomous driving, healthcare, and finance. In the study, this paper first introduces the main algorithms of reinforcement learning; then, it dissects the research progress of reinforcement learning in five fields, such as realizing complex strategy decision-making in the field of game AI, accomplishing high-precision task planning in the field of robotics, realizing scenario-based decision-making in the field of autonomous driving, assisting in personalized treatment planning in the field of healthcare and aiding risk assessment in the field of finance. At the same time, this thesis also discusses the current technical limitations faced by each field and looks forward to possible future development directions. This paper aims to provide comprehensive references and insights for researchers in related fields, and to promote the further development and application of reinforcement learning techniques in various fields

1 Introduction

Reinforcement learning (RL) serves as an important area in the field of artificial intelligence. It is a learning paradigm that learns optimal strategies through continuous trial and error by an intelligent body in a specific environment, and is listed as one of the three mainstream methods of machine learning along with supervised learning and unsupervised learning [1]. The early RL algorithm Q-Learning, which laid a solid theoretical foundation for subsequent research, initially demonstrated the potential of reinforcement learning in solving sequential decision-making problems by continuously optimizing the value function to guide the intelligent body to take action. Meanwhile, Deep Reinforcement Learning (DRL) combines the high-dimensional feature extraction capability of deep learning to address the limitations of traditional RL in complex state spaces.

2022015446@st.cupk.edu.cn

In addition, Reinforcement Learning also efficiently solves complex problems such as robot control, automated driving, intelligent decision-making, etc. with its unique learning method based on a reward mechanism, which provides technical support for the in-depth application of AI technology in various industries. From gaming to industrial automation, from financial investment to medical diagnosis, the application scope of reinforcement learning continues to expand, and its importance has become more and more prominent, and has become a key force to promote artificial intelligence to a higher level. In the gaming field, AlphaGo learns to Go from scratch through self-pairing and reaches the level of surpassing the top human players in a short time, fully demonstrating the powerful ability of reinforcement learning in complex strategy games [2]. In robot control, researchers have enabled robots to learn more natural and efficient ways of movement in simulated environments through reinforcement learning, such as rapid running, flexible grasping, and other complex actions, paving the way for the wide application of robots in the real world. In the field of autonomous driving, reinforcement learning has been used to train the decision-making system of vehicles to make safe and efficient driving decisions autonomously in complex traffic scenarios, which has pushed the pace of autonomous driving technology from the laboratory to real-world applications. These results highlight the excellent ability of reinforcement learning in solving practical problems, and predict that reinforcement learning will realize breakthroughs in more fields, bringing more innovation and change to the development of human society.

This review will summarize the existing classical algorithms for reinforcement learning, and explore and summarize the research results of reinforcement learning in different application domains point out the existing limitations of the algorithms, and finally explore the future research directions.

2 Key Technologies

2.1 Basic Theory

Reinforcement learning addresses the central issue of how intelligence can maximize rewards in complex and uncertain environments. Intelligent bodies use their perception of the state of the environment and the feedback from the environment on their actions to guide them to make better decisions and maximize their rewards. In the process of reinforcement learning, the intelligent body interacts with the environment continuously. Specifically, the intelligent body obtains the current state from the environment and outputs an action based on that state, i.e., makes a decision. This decision is then fed into the environment, which outputs the next state and the reward corresponding to the current decision based on the decision taken by the intelligent body. The goal of intelligence is to obtain as many rewards as possible from the environment.

2.2 Key Algorithms

2.2.1 Proximal Policy Optimization (PPO)

PPO limits the magnitude of strategy update through importance sampling and trimming mechanism to ensure that the difference between the new strategy and the old one is within a controllable range, so as to improve the training stability, as shown in Table 1 below. Meanwhile, PPO adopts the dominance function to reduce the variance. The formula for the trimming objective function is as follows:

$$L^{CLIP}(\theta) = \mathbb{E}_t[min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\, A_t)] \tag{1}$$

Where,$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$, A_t is the dominance function, ϵ is the cropping threshold [3].

2.2.2 Soft Actor-Critic (SAC)

SAC is based on the maximum entropy framework, which maximizes the entropy of the strategies while maximizing the cumulative rewards to enhance exploration, and the application domains are listed in Table 1 below. The algorithm employs two Q-networks (to reduce over-estimation) and one strategy network and balances the objectives by automatically adjusting the entropy weights (temperature parameter α). The soft Q update formula is as follows:

$$Q(s_t, a_t) \leftarrow r_t + \gamma \mathbb{E}_{s_{t+1}}[V(s_{t+1})] \tag{2}$$

Where, $V(s) = \mathbb{E}_{a\sim\pi}[Q(s,a) - \alpha \log \pi(a|s)]$, embodied entropy regularization [4].

2.2.3 Deep Q Network (DQN)

DQN combines Q-learning with deep neural networks, introducing empirical playback (breaks data correlation) and fixed-target networks (stabilizes training). The Q value is updated by minimizing the temporal difference error. The Bellman equation approximation is formulated as follows:

$$L(\theta) = \mathbb{E}\left[\left(r + \gamma \max_{a'} Q_{\text{target}}(s', a') - Q(s,a)\right)^2\right] \tag{3}$$

Where, Q_{target} is the output of the target network [5].

2.2.4 Multi-Agent Deep Deterministic Policy Gradient (MADDPG)

The algorithm uses centralized training and decentralized execution, where the Critic network of each intelligence uses global information (states and actions of other intelligence), while the Actor relies only on local observations. Robustness to policy changes of other intelligences is improved by policy ensemble. Where the deterministic policy gradient update formula is as follows:

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}\left[\nabla_{a_i} Q_i^\pi(s, a_1, \dots, a_N) \nabla_{\theta_i} \pi_i(a_i|s_i)\right] \tag{4}$$

The input of the Critic is the joint action of all agents $(a_1, \dots, a_N)$[6].

2.2.5 Security Reinforcement Learning (SRL)

Limiting the risky behavior of a policy through constrained optimization or risk-sensitive design. The algorithms are characterized in Table 1. commonly found in Constrained Markov Decision Processes (CMDPs). Optimization with constraints:

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \text{s.t.}\, \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t c_t\right] \le C \quad (5)$$

Where, c_t is the cost function, C is the safety threshold [7].

Table 1 Comparison chart of algorithm characteristics

Algorithm	Core features	Main application areas
PPO	Ensure policy update stability by clipping the objective function	Robot control, autonomous driving, financial trading
SAC	Maximum Entropy Framework to Enhance Exploration	Dexterous robot operation, autonomous driving decisions
DQN	Experience Replay + Goal Networks to Resolve Training Dispersion	Game AI, medical image analysis
MADDPG	Centralized Training + Distributed Execution Architecture	Teamwork games such as DOTA2, drone formations
SRL	Formal verification + action constraints	Automated driving, medical dose control

Table 1 compares the characteristics and applications of the five reinforcement learning algorithms. PPO algorithm ensures training stability by limiting the magnitude of policy updates, and is suitable for high-precision scenarios such as robot control and financial trading; SAC enhances the exploration capability based on the entropy-maximization framework, and is commonly used for robot operation and self-driving decision-making; DQN introduces experience playback and goal network mechanism to solve the training dispersion problem, which is mostly used in game AI and medical image analysis; MADDPG adopts centralized training and decentralized execution mode, which is good at dealing with collaborative tasks such as UAV formation; SRL combines formal verification and action constraints, which is prioritized to guarantee the safety of autonomous driving and medical control.

3 Application

3.1 Game AI Field

3.1.1 Complex Strategies and Multi-Intelligence Breakthroughs

In 2013, DeepMind's proposed DQN algorithm was a milestone in this period, the team used DQN to surpass the level of human players in the Atari game with pixel inputs only, which verified the feasibility of deep reinforcement learning with high-dimensional inputs [8]. 2016's AlphaGo series developed by the DeepMind team was a representative achievement, the AI defeated Go world champion Lee Sedol, surpassing the top human player in a complete information game for the first time [9]. In 2019, OpenAI Five used multi-intelligence body collaborative training and course learning to defeat the world champion team in a 5v5 match in Dota 2 by dealing with complex teamwork and long term planning [10], while the DeepMind team developed AlphaStar during the year, on the other hand, beat professional players in StarCraft II [11].

3.1.2 Generalized Capabilities and Open World Exploration

Reinforcement learning techniques have made significant breakthroughs in Model-Based Reinforcement Learning (MBRL), which is characterized by efficient imaginative planning by constructing a model of the world, and is capable of handling more complex open-world environments and multimodal tasks. The DreamerV3 algorithm, proposed in 2022, learns complex skills, such as digging, building, etc., with only a small number of pixel inputs in My World, demonstrating strong generalization ability and adaptability [12]. In 2025, DeepMind proposes an improved visual MBRL framework, Craftax-classic environment, combining the Dyna method, nearest-neighbor tokenizer (NNT), and block teacher forcing (BTF) to outperform human experts in randomly generated 2D open-world level with over 30% improvement in reward efficiency [13]. In addition, in terms of industrial-grade game testing and design, Dead Cells employs RL to generate diverse maps and utilizes reinforcement learning to dynamically adjust level difficulty and style; Blizzard Entertainment deploys AI to simulate a large number of games in Hearthstone Legends to quickly detect an imbalance in card strength. Figure 1 shows the timeline of reinforcement learning in game AI.

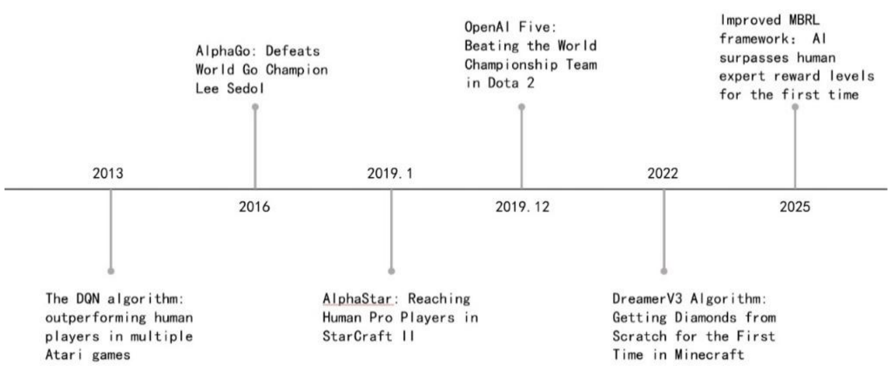

Figure 1 Timeline of Reinforcement Learning in Game AI (Picture credit: Original)

3.2 Robotics Field

The World Model-based Perception (WMP) framework proposed by the ByteDance research team and Shanghai Jiao Tong University enables the Yushu A1 robot to realize difficult actions such as jumping over an 85cm gap by constructing a joint vision-dynamics model, which solves the limitation of traditional privileged learning that relies on manually-designed perceptual features, and significantly improves complex robustness under terrain [14]. Regarding microscale robots, the counterfactual reward mechanism developed by a team from the University of Konstanz, Germany, enables cooperative control of 200 laser-driven micrometer robots for the first time, solves the lazy agent problem, and provides new ideas for minimally invasive healthcare and microchip lab automation [15]. Similarly, a deep reinforcement learning navigation framework for heterogeneous robotic systems improves the task adaptability of logistics and warehousing scenarios through multimodal sensing and cooperative path planning in dynamic environments [16]. For industrial robotic arm control, reinforcement learning strategies based on algorithms such as PPO significantly reduce the cost of traditional manual programming, e.g., in high-precision grasping tasks, autonomous optimization of complex assembly processes is achieved through end-to-end training, which promotes the intelligent upgrading of flexible production lines [17].

3.3 Autonomous Driving Field

In terms of end-to-end system and multi-stage reinforcement learning, the UniAD model proposed by SenseAuto realized the end-to-end fusion of perception and decision-making for the first time, and became the mainstream of the industry by directly generating driving strategies driven by multimodal data [18]. Similarly, the CarPlanner proposed by Zhejiang University and the Rookie Bird team based on reinforcement learning outperforms traditional planning methods for the first time on large-scale real dataset nuPlan, which verifies the efficiency and feasibility of RL in complex dynamic environments [19].

The progress of scenario-based decision-making and simulation training is reflected in the deep optimization of segmented scenarios. For key scenarios such as automatic lane changing, freeway merging and obscured intersections, researchers designed dynamic strategies through reinforcement learning to significantly improve safety and traffic efficiency [20]. Meanwhile, the simulation-driven training framework solves the high-risk problem of real road training, supports large-scale scenario generation and algorithm iteration, and accelerates the model generalization capability improvement [21].

3.4 Medical and Health Field

The application of reinforcement learning in healthcare is mainly focused on two aspects diagnosis and treatment decision-making and drug development. Shravya Shetty's team and DeepMind introduced an AI system for breast cancer screening in 2020, which improved the accuracy and efficiency of cancer screening through RL technology [22]. The MedRAG model proposed by the Nanyang Technological University team in 2025 used knowledge graph inference and reinforcement learning techniques to construct a four-layer fine-grained diagnostic knowledge graph and design an active diagnostic questioning mechanism, which led to an 11.32% improvement in the diagnostic accuracy of chronic pain patients in real clinical data [23].

In terms of drug development, Insilico Medicine's GAN-RL hybrid model optimizes molecule generation paths through reinforcement learning, which improves the efficiency of candidate molecule screening for anticancer drugs by more than three times, significantly accelerating the anticancer drug development cycle [24]. Moreover, the MedRAG model can adjust the questioning strategy based on real-time patient feedback when providing diagnostic suggestions, realizing the dynamic adaptation of diagnosis and treatment plans [23].

3.5 Financial Field

Yonsei University significantly improved the return-risk balance in high volatility markets by integrating a risk constraint mechanism with a deep reinforcement learning framework, achieving a return of 1800% in the cryptocurrency market [25]. In addition, a deep Q-network combining knowledge graph and social sentiment analysis at Kocaeli University improved direction judgment accuracy in Turkish stock market forecasting [26]. Whereas, the Indian Institute of Technology (IIT) research team used dual deep Q-networks for dynamic response mechanism in the Indian stock market and demonstrated strong adaptability to non-stationary characteristics of emerging markets [27].

At the level of technological innovation and model robustness enhancement, to address the problem of noise interference in high-frequency trading, a team from Yonsei University proposed a synergistic framework of wavelet transform and long-short-term memory network (LSTM) to achieve effective feature extraction of financial time series, which improves the stability of the strategy in a noisy environment [28]. On the other hand, the fusion of inverse optimization and deep reinforcement learning constructs a multi-period asset allocation model based on historical preference inference, and its strategy consistently outperforms the S&P 500 benchmark over a five-year cycle [29]. To further address the challenge of training RL models in complex financial environments, a team from Aristotle University of Thessaloniki successfully reduced the volatility variance of strategy gradients in currency trading scenarios by integrating strategy distributions of diversified teacher models through knowledge distillation techniques [30]. The above studies extend the boundaries of RL applications in traditional problems such as order execution and asset pricing, and provide new methodological support for building data-driven intelligent financial decision-making systems.

4 Results

4.1 Technical Limitations

Although reinforcement learning has made significant progress in several fields, its technical limitations still constrain large-scale practical applications, and RL models generally rely on a large amount of trial-and-error training, e.g., in gaming AI, although AlphaGo Zero achieved a level beyond humans by playing against itself, millions of simulations were required, and the training cost was extremely high [31]. The robotics field also faces the problem of insufficient generalization ability, and models trained in specific environments are difficult to migrate to real-world complex scenarios, e.g., industrial robots need to repeatedly adjust parameters to adapt to different production line requirements. In addition, in healthcare, RL's dependence on high-quality labeled data limits its application in disease prediction and drug discovery, a challenge that is further exacerbated by the privacy and incompleteness of medical data.

Another core limitation is dynamic environment adaptation and interpretability. In automated driving, RL needs to deal with changing road conditions and emergencies in real time, but the reliability of existing models in extreme weather or complex interaction scenarios is still questionable. Although high-frequency trading systems in the financial field can optimize their strategies through RL, rapid changes in the market environment often lead to model obsolescence, and the “black-box” nature makes it difficult for regulators to trust the decision-making logic.

4.2 Future Outlook

The future development of reinforcement learning will center around multimodal fusion and simulation breakthroughs. In the field of autonomous driving, a more robust perception-decision integration model can be constructed by combining vision, radar and LIDAR data, and a high-precision simulation environment can be utilized to accelerate training and reduce the risk of real vehicles. In the medical field, federated learning and multi-center data collaboration are expected to solve the problem of data silos, and personalized treatment recommendation systems combined with reinforcement learning will serve patients more accurately.

On the other hand, interpretability enhancement and dynamic adaptation optimization will become the key direction. Through the introduction of Neuro-Symbolic AI, the decision-making process of RL models can be decomposed into logic rules and data-driven modules, improving the transparency of medical diagnosis and financial risk control. Meanwhile, Meta-Learning and online learning techniques will help models adapt to environmental changes in real time, for example, financial trading systems can respond to market fluctuations by dynamically adjusting reward functions, and robots can quickly adapt to new tasks with a small amount of demonstration data.

5 Conclusions

This paper provides a systematic review of the cross-domain applications of reinforcement learning technology, focusing on analyzing its practical progress in five major fields: game AI, robot control, autonomous driving, healthcare and financial decision-making. The significance of this paper lies in the fact that through the compilation of multi-disciplinary application maps, it reveals the universal value of reinforcement learning as a generalized decision-making framework, and its characteristics of interacting with the environment and evolving autonomously fit the core demand for dynamic decision-making systems in the era of intelligence; Second, the comparative analysis of cross-disciplinary application cases shows that despite the significant differences in state space construction and reward function design in different domains, all of them are able to achieve optimization effects that are difficult to achieve by traditional methods through the reinforcement learning framework. It is worth noting that with the emergence of new paradigms such as multi-intelligent body reinforcement learning and meta-learning, this technique shows more potential in the direction of distributed system optimization and cross-task migration. Future research needs to focus on key technological breakthroughs such as algorithm interpretability, sample efficiency improvement, and security constraint reinforcement, while deepening the integration and innovation with vertical domain knowledge to promote the application of reinforcement learning in emerging scenarios such as energy scheduling and education personalization.

References

1. Bai Tian, Lu Yao, Li Chu, and He Jialiang, "Game intelligent guidance algorithm based on deep reinforcement learning," Journal of Jilin University (Science Edition), vol. 63, no. 1, pp. 91-98, 2025, doi: 10.13413/j.cnki.jdxblxb.2023555.
2. D. Silver, et al., "Mastering the game of Go with deep neural networks and tree search," Nature, vol. 529, no. 7587, pp. 484-489, 2016.
3. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.
4. H. Yong, J. Seo, J. Kim, M. Kim, and J. Choi, "Suspension control strategies using switched soft actor-critic models for real roads," IEEE Trans. Ind. Electron., vol. 70, no. 1, pp. 824-832, 2022.
5. V. Mnih, et al., "Human-level control through deep reinforcement learning," Nature, vol. 518, no. 7540, pp. 529-533, 2015.
6. R. Lowe, et al., "Multi-agent actor-critic for mixed cooperative-competitive environments," Adv. Neural Inf. Process. Syst., vol. 30, 2017.
7. J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in Proc. Int. Conf. Mach. Learn., 2017, pp. 22-31.
8. V. Mnih, et al., "Playing Atari with deep reinforcement learning," arXiv preprint arXiv:1312.5602, 2013.

9. D. Silver, et al., "Mastering the game of Go with deep neural networks and tree search," Nature, vol. 529, no. 7587, pp. 484-489, 2016.
10. C. Berner, et al., "Dota 2 with large-scale deep reinforcement learning," arXiv preprint arXiv:1912.06680, 2019.
11. O. Vinyals, et al., "Grandmaster level in StarCraft II using multi-agent reinforcement learning," Nature, vol. 575, no. 7782, pp. 350-354, 2019.
12. D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," arXiv preprint arXiv:2301.04104, 2023.
13. A. Dedieu, et al., "Improving Transformer World Models for Data-Efficient RL," arXiv preprint arXiv:2502.01591, 2025.
14. H. Lai, et al., "World Model-based Perception for Visual Legged Locomotion," arXiv preprint arXiv:2409.16784, 2024.
15. V. L. Heuthe, E. Panizon, H. Gu, and C. Bechinger, "Counterfactual rewards promote collective transport using individually controlled swarm microrobots," Sci. Robot., vol. 9, no. 97, eado5888, 2024.
16. Y. Chen and J. Xiao, "Target Search and Navigation in Heterogeneous Robot Systems with Deep Reinforcement Learning," Mach. Intell. Res., vol. 22, no. 1, pp. 79-90, 2025.
17. T. Miki, et al., "Learning robust perceptive locomotion for quadrupedal robots in the wild," Sci. Robot., vol. 7, no. 62, eabk2822, 2022.
18. Y. Hu, et al., "Planning-oriented autonomous driving," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 17853-17862.
19. D. Zhang, et al., "CarPlanner: Consistent Auto-regressive Trajectory Planning for Large-scale Reinforcement Learning in Autonomous Driving," arXiv preprint arXiv:2502.19908, 2025.
20. P. Wang, C. Y. Chan, and A. de La Fortelle, "A reinforcement learning based approach for automated lane change maneuvers," in Proc. IEEE Intell. Vehicles Symp. , 2018, pp. 1379-1384.
21. C. Wu, et al., "Flow: Architecture and Benchmarking for Reinforcement Learning in Traffic Control," arXiv preprint arXiv:1710.05465, 2017.
22. S. M. McKinney, et al., "International evaluation of an AI system for breast cancer screening," Nature, vol. 577, no. 7788, pp. 89-94, 2020.
23. X. Zhao, S. Liu, S. Y. Yang, and C. Miao, "MedRAG: Enhancing Retrieval-augmented Generation with Knowledge Graph-Elicited Reasoning for Healthcare Copilot," arXiv preprint arXiv:2502.04413, 2025.
24. F. Ren, et al., "A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models," Nat. Biotechnol. , vol. 43, no. 1, pp. 63-75, 2025.
25. W. Shin, S. J. Bu, and S. B. Cho, "Automatic financial trading agent for low-risk portfolio management using deep reinforcement learning," arXiv preprint arXiv:1909.03278, 2019.
26. A. B. Altuner and Z. H. Kilimci, "A novel deep reinforcement learning based stock direction prediction using knowledge graph and community aware sentiments," arXiv preprint arXiv:2107.00931, 2021.
27. S. Bajpai, "Application of deep reinforcement learning for Indian stock trading automation," arXiv preprint arXiv:2106.16088, 2021.

28. J. Lee, H. Koh, and H. J. Choe, "Learning to trade in financial time series using high-frequency through wavelet transformation and deep reinforcement learning," Appl. Intell. , vol. 51, no. 8, pp. 6202-6223, 2021.
29. H. Wang and S. Yu, "Robo-advising: Enhancing investment with inverse optimization and deep reinforcement learning," in Proc. IEEE Int. Conf. Mach. Learn. Appl., 2021, pp. 365-372.
30. A. Tsantekidis, N. Passalis, and A. Tefas, "Diversity-driven knowledge distillation for financial trading using deep reinforcement learning," Neural Netw., vol. 140, pp. 193-202, 2021.
31. D. Silver, et al., "Mastering the game of Go without human knowledge," Nature, vol. 550, no. 7676, pp. 354-359, 2017.