Proceedings of the 4th International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2026.35058

Intelligent Decision-Making Optimization Based on Deep Reinforcement Learning

Haihang Gu

School of Economics and Management, Xidian University, Xi'an, China
24061300116@stu.xidian.edu.cn

Abstract. In recent years, deep reinforcement learning has seen notable advances in areas such as game playing and autonomous driving. How the underlying reasoning and decision-making processes actually develop and improve remains an open question that deserves closer attention. This paper starts from the most basic decision-making dilemma and traces the optimization path of deep reinforcement learning in reasoning and decision-making along the logical thread of sequential choice under complexity. Through a comparative look at classic strategies in the multi-armed bandit problem, the theoretical structure of Markov decision processes, and major deep reinforcement learning algorithms from the past five years, the paper shows that the heart of deep reinforcement learning optimization lies in a shift from model-based deduction toward interaction-based experience building. The results suggest that current algorithms continue to grapple with a trade-off between sample efficiency and generalization, while structural causal reasoning is drawing growing interest as a new research direction. The aim of this work is to offer a reference point for thinking about the nature of decision-making and for designing more effective learning algorithms.

Keywords: deep reinforcement learning, reasoning and decision-making, multi-armed bandit, exploration strategies

1. Introduction

Deep reinforcement learning is quietly altering the landscape of machine decision-making, and evidence of this shift now appears across a remarkably broad spectrum of domains. The most visible landmark remains AlphaGo's victory over a human world champion in Go, yet equally telling are the less theatrical but more pervasive deployments: autonomous vehicles threading their way through disordered intersections, or recommender systems that seem to sense a user's next interest before it is fully formed. As the core objective of deep reinforcement learning, intelligent decision-making aims to make agents to continuously optimize their behavioral strategies through interaction and learning in dynamic and uncertain environments, finally maximizing long-term returns. Different from the traditional decision-making methods which rely on predefined rules or static optimization models, deep reinforcement learning gives agents the ability to self-evolve through continuous trial and error. The agent does not need to know the complete operational mechanism of the environment in advance. Instead, through repeated interactions with the environment, it gradually accumulates

DOI: 10.54254/2753-8818/2026.35058

experience and gains valuable behavioral patterns from sparse reward signals. This paradigm of "learning from interaction" makes intelligent decision-making to have adaptability and robustness.

This paper focuses on the key aspects of deep reinforcement learning in decision optimization, systematically reviewing its development trajectory and core mechanisms. Starting from the classic multi-armed bandit problem, this paper analyzes the fundamental dilemma of exploration and exploitation in uncertain environments. Then this paper introduce the concept of state transition and discusses the theoretical framework of Markov decision processes and their limitations in dynamic programming. Finally, this paper reviews the major advancements in deep reinforcement learning and the breakthrough improvements in reasoning efficiency and decision-making quality over the past five years. Through this review, this paper aims to present the internal logic of deep reinforcement learning in decision optimization and provide insights that can inform future research.

2. Reasoning in classical decision-making models

2.1. The multi-armed bandit problem

The multi-armed bandit problem makes a fitting entry point for reinforcement learning, because it captures the essence of decision-making. The user can never be sure whether an untried option is better than a known good one. In the past five years, this familiar problem has been pushed into more challenging territory. Xie and colleagues examined how to estimate rewards robustly when the underlying distributions grow heavy-tailed. They proposed a quantile-based adjustment to the Upper Confidence Bound algorithm and showed that it delivers regret bounds noticeably tighter than earlier formulations [1]. That line of work carries a wider lesson—relax a few tidy distributional assumptions, and even problems long considered closed can sprout new theoretical questions.

Among the many strategies devised for the bandit problem, Thompson sampling continues to draw interest, largely because of the distinctive way it handles uncertainty. Each arm is assigned a distribution that reflects belief about its true reward probability, and the action taken is simply a draw from those distributions. Uncertainty, in other words, is not something to be averaged out before acting; it is woven directly into the mechanism that selects the next move. The choice of arm follows not from a settled conviction about which is best, but from the fog of doubt that remains.

When the setting expands to include multiple agents, Thompson sampling still has its strengths, though the complications multiply. Taïga and colleagues observed that when agents sample independently and at the same time, their individual exploratory steps can start colliding, dragging down the system's overall efficiency. Their answer was a coordinated variant: agents share what they know about uncertainty, so the collective exploration becomes less disjointed and more purposeful [2]. That habit of folding uncertainty directly into decision-making has carried forward into more recent deep reinforcement learning work. Some algorithms, for example, inject noise straight into network parameters, giving the learned policy randomness to achieve a more organic kind of exploration.

2.2. From single-step to sequential decision-making

The multi-armed bandit problem characterizes decision-making in static settings, but the choices often unfold sequentially in real world, which is a vital problem that lies at the center of Markov decision processes (MDP). MDP describes an interactive mechanism: an agent situated in a given state takes an action, then the environment responds with an immediate reward and the agent transitions to a subsequent state. The agent's objective is to maximize long-term cumulative reward.

DOI: 10.54254/2753-8818/2026.35058

As a classical solution method for MDP, dynamic programming rests on a core recursive insight: the value of the current state equals the sum of the immediate reward and the discounted value of future states. Bellman first formalized this recursive relationship, which remains foundational to the theory of reinforcement learning.

However, dynamic programming also carries a significant limitation. It presupposes access to a complete model of the environment which is the transition states and reward values corresponding to each action taken in each state. This requirement is almost impossible to satisfy in real-world scenarios. This paper argues that the limitations of dynamic programming expose a fundamental predicament in classical decision theory. It relies on the cognitive model of first mastering the rules of how the world works, then making optimal decisions. In contrast, real-world decision-making frequently follows an inverted way. The agent acquires knowledge of environmental regularities through sustained interaction and adjusts its policy continuously. This characteristic marks the essential departure of reinforcement learning from dynamic programming and it constitutes the underlying reason why deep reinforcement learning has been able to achieve significant advances.

3. Optimization in deep reinforcement learning

3.1. The evolution of exploration strategies

Although traditional exploration strategies such as ε-greedy which selects a random action with a fixed small probability are simple to implement, it is inefficient in high-dimensional state and action spaces. Researchers have gained inspiration from the multi-armed bandit problem in recent years, which extend the core logic of Thompson sampling into deep reinforcement learning settings. One approach involves introducing randomness in the parameter space rather than the action space. The parameter noise method proposed by García Fernández and his colleagues represent an early effort in this direction. However, it is relatively sensitive to the noise scale in practical applications. Recently, an adaptive parameter noise algorithm has demonstrated better performance over the original parameter noise approach on continuous control tasks through adjusting the noise variance online [3].

Another solution obtained from research on intrinsic motivation. Curiosity-driven exploration stands as a seminal contribution in this area, but subsequent investigations have identified a shortcoming. In certain environments, the agent may become addicted to "television noise"—states that are highly random but lack semantic meaning. To overcome this shortcoming, Sun proposed an intrinsic reward mechanism based on prediction uncertainty. Use an ensemble network to estimate prediction uncertainty and provide exploration rewards only when the uncertainty stems from a lack of knowledge rather than the environmental noise [4]. Experimental results show that this method achieved better results than traditional curiosity-driven approaches in maze environments containing random elements.

3.2. Efficiency optimization

Efficiency optimization focuses on achieving faster policy convergence under a limited budget of interaction samples. Parameter noise methods bring perturbations in the parameter space to enable continuous policy exploration while maintaining action smoothness, then avoiding oscillations caused by random exploration in the action space. Through adjusting the noise variance online, the adaptive parameter noise algorithm maintains high randomness in the early stage of exploration to cover a wide state space and automatically reduces the noise intensity when the policy approaches

DOI: 10.54254/2753-8818/2026.35058

convergence. This mechanism establishes a dynamic balance between exploration and exploitation. These optimizations improve sample efficiency and reduce the number of environmental interactions while guaranteeing the quality of the final policy.

3.3. Agent optimization

Agent optimization focuses on the structure and intention of exploration behaviors themselves, pursuing a more intelligent decision-making mechanism. Instead of depending on random perturbations supplied by the external environment, intrinsic motivation methods endow the agent with the capacity to generate exploratory objectives autonomously. Curiosity-driven exploration makes the agent to choose the states that promise "cognitive progress" to make the agent not only ventures into unknown territory but also discerns which forms of ignorance merit active investigation. This optimization changes the agent from a state machine that passively responds to environmental feedback into an autonomous system with self-regulatory capabilities.

3.4. Constrained optimization

Lagrangian methods and primal-dual optimization are the principal technical approaches within this domain. The basic idea is incorporating the constraints into the objective function which weighted by Lagrange multipliers and then converting the constrained optimization problem into an unconstrained minimax problem. During iterative updates, the agent adjusts both the policy parameters and the Lagrange multipliers to seek an equilibrium between reward maximization and constraint satisfaction [5]. On this basis, constrained policy adjustment methods provide a more direct mechanism for constraint satisfaction. They enforce constraints explicitly at each step of policy updates to ensure that the policy generated at every iteration remains are in the feasible region. This makes them suitable for application scenarios with extremely high safety requirements [6].

4. Bottlenecks in current research and future directions

4.1. The sample efficiency bottleneck

The most advanced algorithms still need millions of environmental interactions to acquire competence in a single task. But humans often need only a few attempts to study a new skill. The reason of this gap is that current algorithms are short of the capacity of causal reasoning. Cao and colleagues introduced a framework that incorporates structural causal models into reinforcement learning to enable the agent to learn the causal structure governing action–outcome relationships concurrently with policy acquisition [7]. They confirmed that under certain identifiability conditions, the learned causal structure yields substantial improvements in policy performance under out-of-distribution tests.

4.2. The generalization capacity bottleneck

Another issue is generalization. An agent trained in a specific environment often finds itself helpless when transferred to a slightly different environment. A systematic survey of the generalization problem has been provided by Kirk and colleagues. They categorize the generalization problem into three types: intra-task generalization, inter-task generalization and environment dynamics

DOI: 10.54254/2753-8818/2026.35058

generalization. They also point out that current algorithms rely excessively on the statistical regularities of the training environment, failing to learn truly transferable decision rules [8].

4.3. Potential future directions

Based on the analysis, this paper suggests that deep reinforcement learning may evolve in the following directions with respect to reasoning and decision-making in the coming years. First, the integration of causal reasoning with reinforcement learning stands as a promising direction. The causal understanding of the environment facilitates rapid adaptation to novel settings. Second, the convergence of large language models and reinforcement learning merits close attention. Carta attempted to use a language model as the "brain" of the reinforcement learning system to handle high-level planning tasks, while leaving low-level control to traditional reinforcement learning algorithms [9]. In related work, Rowe and colleagues proposed a language model to generate possible successful trajectories and then used to pretrain a policy network, followed by fine-tuning in the real environment [10]. Third, the development of more efficient exploration strategies remains a pressing objective. Drawing inspiration from human to leverage prior knowledge to guide exploration rather than commencing from scratch with undirected trial and error.

5. Conclusion

The optimization of deep reinforcement learning is primarily reflected at two levels. First, it embodies a transition from deduction to experiential accumulation. Traditional dynamic programming presupposes a fully specified model of the environment and arrives at optimal solutions through mathematical computation. But deep reinforcement learning accumulates experience through environmental interaction and derives decision rules directly from data. This shift makes the application of reinforcement learning to complex problems which the model is unknown. Second, exploration strategies have become markedly more intelligent and proactive. Starting from rudimentary random exploration, progressing through uncertainty-based exploration, and culminating in exploration driven by intrinsic motivation such as curiosity, the exploration mechanisms within deep reinforcement learning increasingly approximate the cognitive patterns observed in human learning.

Current research still faces lots of considerable challenges. Issues such as low sample efficiency and the intricacy of multi-agent interactions need to be solved. Encouragingly, from the multi-armed bandit to deep reinforcement learning, the path of decision theory has evolved from simplicity to complexity and from abstraction to concreteness. Future intelligent agents are expected not merely to make decisions, but to comprehend the causal logic underlying those decisions; not only to function within familiar environments, but to explore and adapt amid the unknown.

References

[1] Xie, H., Tang, Q., & Zhu, Q. (2023). A multiplier bootstrap approach to designing robust algorithms for contextual bandits. IEEE Transactions on Neural Networks and Learning Systems, 34(12), 9887–9899.
[2] Taïga, A. A., Courville, A. C., & Bellemare, M. G. (2022). Introducing coordination in concurrent reinforcement learning. Paper presented at the International Conference on Learning Representations.
[3] García Fernández, J., Ahmad, N., & van Gerven, M. (2024). Ornstein-Uhlenbeck adaptation as a mechanism for learning in brains and machines. Entropy, 26(12), 1125.
[4] Sun, C. (2023). Curiosity-driven learning in artificial intelligence and its applications [Doctoral dissertation, Nanyang Technological University].

DOI: 10.54254/2753-8818/2026.35058

[5] Bai, Q., Bedi, A. S., Agarwal, M., et al. (2022). Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In Proceedings of the AAAI Conference on Artificial Intelligence, 36(4), 3682–3689.
[6] Liu Z, Cen Z, Isenbaev V, et al. Constrained variational policy optimization for safe reinforcement learning [C]//International Conference on Machine Learning. PMLR, 2022: 13644-13668.
[7] Cao, H., Feng, F., Fang, M., Dong, S., Yang, T., Huo, J., & Gao, Y. (2025). Towards empowerment gain through causal structure learning in model-based reinforcement learning. Paper presented at the Thirteenth International Conference on Learning Representations, Singapore.
[8] Kirk, R., Zhang, A., Grefenstette, E., & Rocktäschel, T. (2023). A survey of generalisation in deep reinforcement learning. Journal of Artificial Intelligence Research, 78, 1–59.
[9] Carta, T., Romac, C., Wolf, T., et al. (2023). Grounding large language models in interactive environments with online reinforcement learning. In International Conference on Machine Learning, 3679–3710.
[10] Rowe, L., de Schaetzen, R., Girgis, R., Pal, C., & Paull, L. (2025). Poutine: Vision-language-trajectory pre-training and reinforcement learning post-training enable robust end-to-end autonomous driving. arXiv preprint, arXiv: 2506.11234.