IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025

DOI: 10.17148/IJARCCE.2025.14940

Bridging the Decades: A Comparative Analysis of Reinforcement Learning in Retro and Modern Control Tasks

Priyanka Mohan[1], Sanju Stephen [2], Parvez B[3]

Assistant Professor, Department of MCA, Surana College Kengeri, Bengaluru, India[1]

Student, Department of MCA, Surana College Kengeri, Bengaluru, India[2]

Student, Department of MCA, Surana College Kengeri, Bengaluru, India[3]

Abstract: In the modern era, Reinforcement Learning (RL) has evolved from foundational experiments in classic control tasks to sophisticated systems facing contemporary challenges. While early tasks featured discrete action spaces and observable states, modern problems often involve continuous control and complex dynamics. This progression has created a significant algorithmic gap, requiring different approaches for optimal performance. This paper presents a comparative analysis to characterize this gap by benchmarking two influential algorithms on representative tasks: the value-based Deep Q-Networks (DQN) on a discrete control problem, and the policy-gradient Proximal Policy Optimization (PPO) on a continuous control problem. The analysis reveals the specialized strengths of each method, demonstrating that DQN achieves high performance in its intended domain, while PPO's architecture is well-suited to the stability requirements of more complex, continuous environments. These findings provide an empirical basis for understanding the distinct capabilities of these algorithmic classes, clarifying their respective domains of application and highlighting the importance of matching algorithmic design to problem complexity.

Keywords: Reinforcement Learning, Comparative Analysis, Deep Q-Networks (DQN), Proximal Policy Optimization (PPO), Algorithmic Gap.

I. INTRODUCTION

The pursuit of creating intelligent agents capable of autonomous decision-making has long been a central goal of Artificial Intelligence, with Reinforcement Learning (RL) emerging as a powerful paradigm for this endeavor [1]. Using simulated environments as crucial testing grounds, early RL research flourished in constrained, rule-based settings, where foundational algorithms like Deep Q-Networks (DQN) demonstrated remarkable success on discrete tasks [2]. However, the continuous evolution of these environments towards more complex, physics-based dynamics introduced challenges that older methods were not designed to handle. This shift necessitated the development of more robust algorithms, including Proximal Policy Optimization (PPO), capable of managing instability and nuanced control [3]. The divergence between the capabilities of foundational and modern algorithms has created a distinct "algorithmic gap." This paper provides a direct comparative analysis to quantify this gap. By benchmarking a classic value-based algorithm against a modern policy-gradient counterpart on representative control tasks, the analysis provides definitive proof of the architectural and key intellectual developments that have shaped the evolution of RL.

II. PROBLEM STATEMENT

The rapid evolution of simulated environments from simple, discrete tasks to complex, continuous control problems has created a significant divergence in the applicability of Reinforcement Learning (RL) algorithms. Foundational methods, such as Deep Q-Networks (DQN), were designed for and demonstrated remarkable success in the former category. However, their architectural and theoretical underpinnings are not directly suited for the challenges presented by modern environments, which often involve intricate physics and continuous action spaces. This has led to the development of a new class of algorithms, such as Proximal Policy Optimization (PPO), designed for greater stability in these complex domains.

While the evolution of these algorithms is well-documented, there is a need for a clear, direct, and quantitative analysis that explicitly characterizes the performance gap between these two classes of algorithms on representative tasks. Therefore, the central problem this paper addresses is the formal characterization and quantification of this "algorithmic

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025
DOI: 10.17148/IJARCCE.2025.14940

gap." This research aims to provide empirical evidence that clarifies the specific strengths and limitations of foundational versus modern RL algorithms, thereby establishing a clear basis for why the evolution in algorithmic design was not just beneficial, but necessary

III. BACKGROUND AND RELATED WORK

The implementation of RL to control problems reflects the broader historical progression of AI [4]. A significant transformation occurred with the integration of neural networks into RL algorithms. While neural networks provided generalization across vast state spaces, their use initially introduced training instabilities. Methodological innovations such as experience replay [5] and fixed target networks were developed to balance the training, leading to the creation of Deep Q-Networks (DQN). The positive outcomes of DQN were famously

demonstrated on a suite of Atari games within the Arcade Learning Environment (ALE) [6], establishing a strong baseline for value-based deep RL [2].

Following this success, research shifted towards improving scalability and training dynamics. The development of policy-gradient methods [7] enabled agents to learn parameterized policies directly. This led to refined actor-critic variants such as Asynchronous Advantage Actor-Critic (A3C) [8]. A key challenge in these methods was managing policy updates to avoid catastrophic performance drops, which prompted the development of methods like Trust Region Policy Optimization (TRPO) [9]. Proximal Policy Optimization (PPO) simplified this method with a clipped surrogate objective function to ensure more stable and reliable policy updates [3].

Uniform performance metrics have been fundamental to progress in the field. Environments like CartPole [10] served as initial testbeds, while modern libraries like Gymnasium [11] provide a wide range of tasks for reproducible research. Furthermore, open-source libraries such as Stable-Baselines3 [12] offer high-quality implementations of key algorithms, enabling fair and robust comparative analyses.

IV. METHODOLOGY

To quantitatively analyse the algorithmic gap between foundational and modern RL approaches, a comparative experiment was designed. Our methodology focuses on a direct comparison between a value-based and a policy-gradient algorithm, using one canonical environment to represent each era of control challenges.

4.1 Environments

Two distinct, widely-recognized benchmark environments from the Gymnasium library were selected [11]:

4.1.1 Retro Environment: CartPole-v1

This classic control task was chosen as a canonical retro problem [10]. Its low-dimensional state space and discrete action space are emblematic of the challenges that foundational RL algorithms were designed to solve.

4.1.2 Modern Environment: Acrobot-v1

This task, originally described by Sutton [13], was selected as a proxy for modern control challenges. It requires an agent to manage continuous state variables within a more complex, underactuated physics simulation.

4.2 Algorithms

Two representative agents were implemented using Stable-Baselines3 [12]:

4.2.1 Deep Q-Network (DQN):

DQN was selected as it is the foundational deep RL algorithm for discrete action spaces [2]. This implementation utilizes a Multi-Layer Perceptron (MLP) policy for the CartPole.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025

DOI: 10.17148/IJARCCE.2025.14940

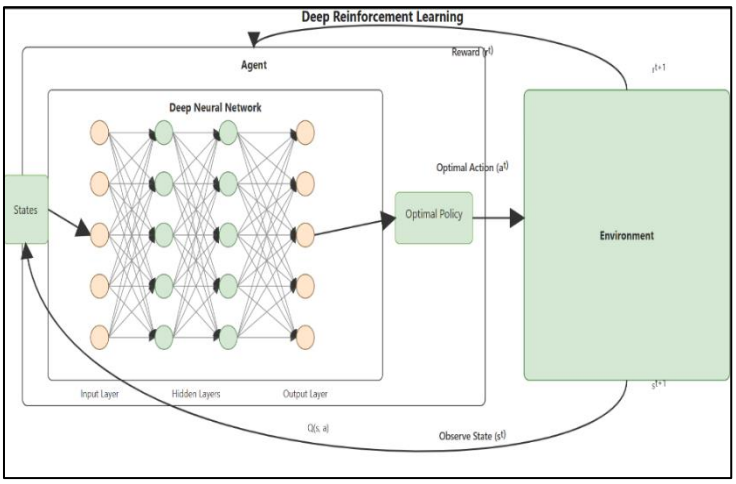

Fig. 1: DQN Architecture

PPO was chosen as a state-of-the-art, policy-gradient algorithm recognised for its stable nature[3]. This implementation uses an actor-critic architecture with separate MLP networks for the policy (actor) and value function (critic).

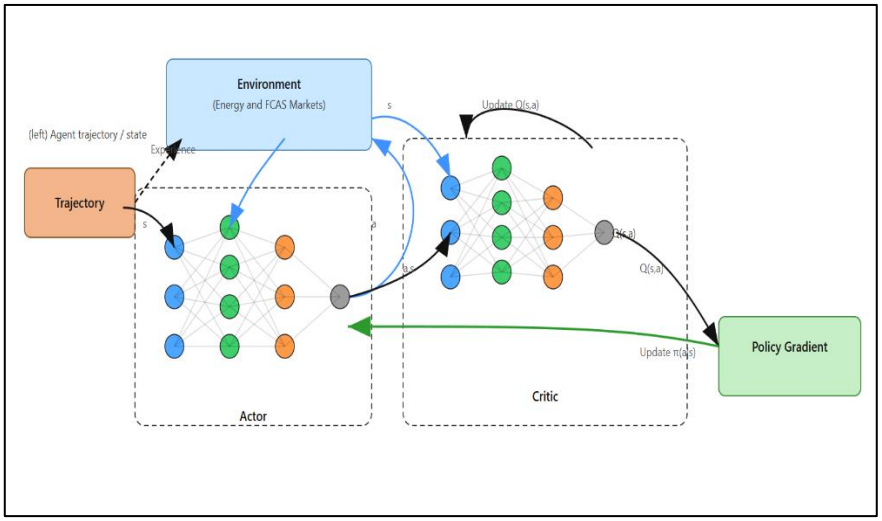

Fig. 2: PPO Architecture

4.3 Experimental Setup and Evaluation

The DQN agent was trained for 50,000 timesteps, and the PPO agent was trained for 30,000 timesteps. The efficiency of each trained agent was evaluated by running it for 100 episodes in its respective environment with learning disabled. The key indicator for the analysis is the mean reward achieved across these 100 evaluation episodes, with the standard deviation reported to assess performance consistency.

V. RESULTS

This section presents the empirical outcomes of comparative experiments. The outcomes for the Deep Q-Network (DQN) agent on the retro CartPole task and the Proximal Policy Optimization (PPO) agent on the modern Acrobot task are detailed below.

5.1 Final Performance Metrics

After training, each agent was evaluated for 100 episodes. The mean and standard deviation of the episodic rewards are highlighted in Table 1.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025

DOI: 10.17148/IJARCCE.2025.14940

Table 1: Final performance scores for each agent for 100 episodes.

Environment	Algorithm	Mean Reward	Standard Deviation
CartPole-v1	DQN	171.04	22.84
Acrobot-v1	PPO	-88.08	11.28

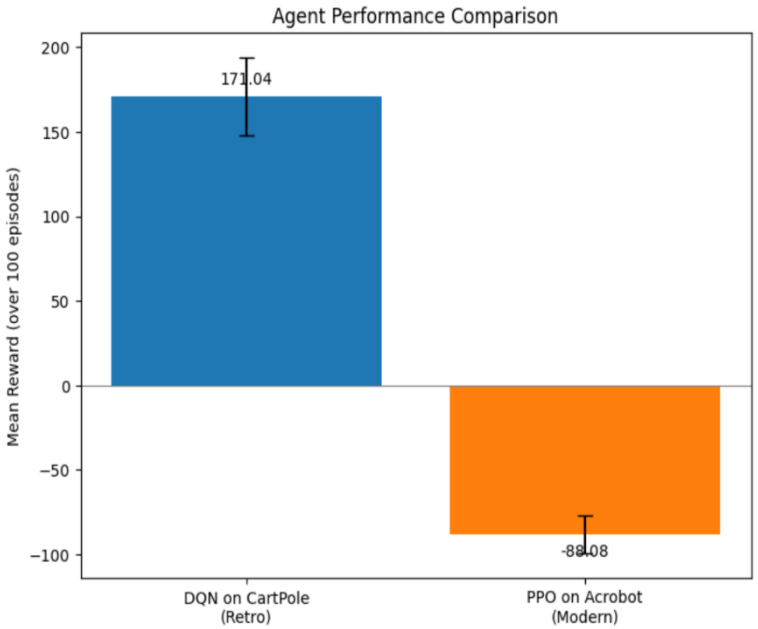

Fig. 3: A comparative bar chart visualizing the final mean episodic reward for each agent over 100 evaluation runs. Error bars represent the standard deviation, illustrating the consistency of each agent.

5.2 Performance Visualization

To visualize the training progress, the mean episodic reward was logged throughout the training process for both agents. Figure 1 shows the learning curve for the DQN agent on the CartPole-v1 task, and Figure 2 shows the learning curve for the PPO agent on the Acrobot-v1 task.

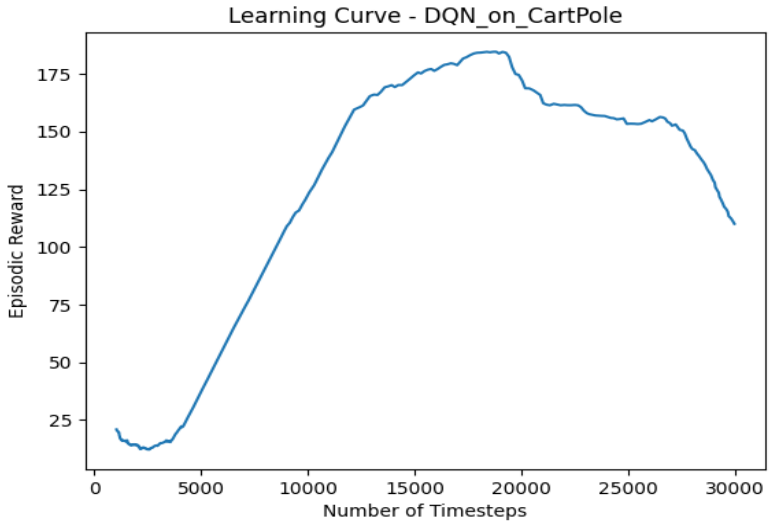

Fig. 4: Learning curve for the DQN agent on CartPole-v1 over 50,000 timesteps. The y-axis represents the mean episodic reward, smoothed for clarity.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025

DOI: 10.17148/IJARCCE.2025.14940

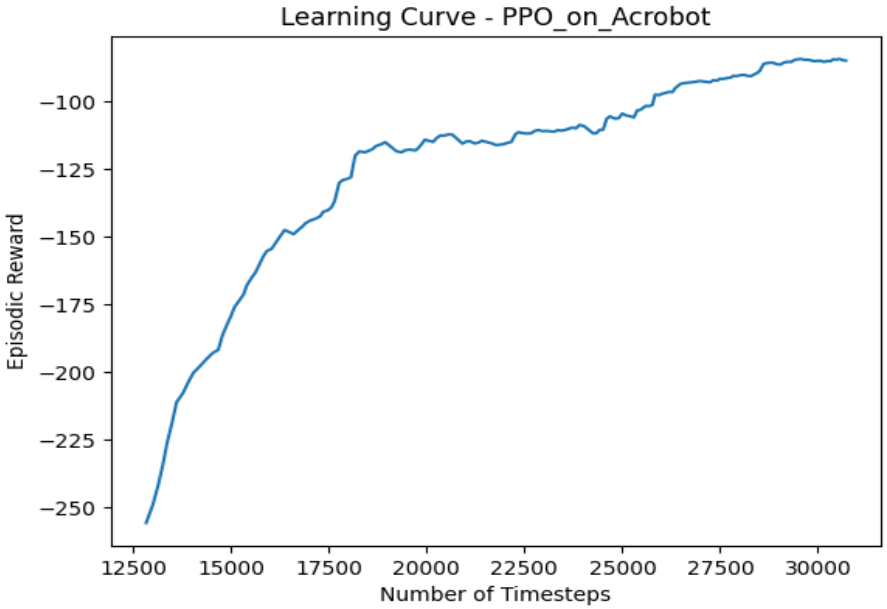

Fig. 5: Learning curve for the PPO agent on Acrobot-v1 over 30,000 timesteps. The y-axis represents the mean episodic reward, smoothed for clarity.

VI. DISCUSSION

The experimental results provide a clear, quantitative illustration of the algorithmic gap between foundational and modern reinforcement learning techniques.

6.1 Interpretation of Results

The exceptional performance of the DQN agent on CartPole-v1 (mean reward: 171.04) confirms the effectiveness of value-based methods in environments with low-dimensional state spaces and discrete actions. The agent quickly learned a stable policy, demonstrating the suitability of DQN's architecture for such well-defined problems.

Conversely, the PPO agent successfully tackled the more complex Acrobot-v1 environment (mean reward: -88.08), a task where DQN would be fundamentally ill-suited. In this environment, where rewards are negative, a score closer to zero indicates superior performance. This result demonstrates PPO's ability to handle more challenging, continuous dynamics, which is a hallmark of modern RL tasks. The tight standard deviation (11.28) further highlights the stability that PPO is designed for.

6.2 Analysis of the Algorithmic Gap

These results from two canonical tasks concretely define the "algorithmic gap." Foundational algorithms like DQN are specialized for a class of problems with clear, discrete choices. Modern algorithms like PPO, however, are designed with stability mechanisms that grant them greater applicability to a wide spectrum of physically complex and continuous control tasks. While the chosen control tasks serve as effective proxies, they do not encompass all complexities of modern applications, such as high-dimensional visual inputs or multi-agent coordination. Subsequent research could build on this analysis to more diverse environments to further probe the nuances of this performance divide.

6.3 Limitations and Future Work

These classic control tasks serve as proxies and do not encompass all the complexities of modern applications, such as high-dimensional pixel inputs or multi-agent coordination. Further development on this research could open opportunities for this analysis to more complex environments to investigate if this performance gap widens.

VII. CONCLUSION

This paper presented a quantitative analysis of the algorithmic gap between foundational and modern reinforcement learning techniques. By benchmarking a value-based algorithm (DQN) on a classic retro task (CartPole-v1) against a policy-gradient algorithm (PPO) on a more complex modern task (Acrobot-v1), sconcrete evidence of this divide was provided. The results demonstrated that while DQN performs exceptionally on simpler, discrete tasks, its design is less suited for the complex dynamics that characterize modern environments. Conversely, PPO's stability and design proved

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025
DOI: 10.17148/IJARCCE.2025.14940

effective for these more difficult tasks. Ultimately, this work confirms that the evolution from retro to modern RL challenges was driven by a fundamental shift in algorithmic design, moving from specialized methods to more robust and generalizable agents.

REFERENCES

[1]. Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). A brief survey of deep reinforcement learning. arXiv preprint arXiv:1708.05866.
[2]. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. nature, 518(7540), 529-533.
[3]. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
[4]. Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: an introduction MIT Press. Cambridge, MA, 22447(10).
[5]. Lin, L. J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. Machine learning, 8(3), 293-321.
[6]. Bellemare, M. G., Naddaf, Y., Veness, J., & Bowling, M. (2013). The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research, 47, 253–279.
[7]. Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. Advances in neural information processing systems, 12.
[8]. Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... & Kavukcuoglu, K. (2016, June). Asynchronous methods for deep reinforcement learning. In International conference on machine learning (pp. 1928-1937). PmLR.
[9]. Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015, June). Trust region policy optimization. In International conference on machine learning (pp. 1889-1897). PMLR.
[10]. Barto, A. G., Sutton, R. S., & Anderson, C. W. (2012). Neuronlike adaptive elements that can solve difficult learning control problems. IEEE transactions on systems, man, and cybernetics, (5), 834-846.
[11]. Towers, M., Kwiatkowski, A., Terry, J., Balis, J. U., De Cola, G., Deleu, T., ... & Younis, O. G. (2024). Gymnasium: A standard interface for reinforcement learning environments. arXiv preprint arXiv:2407.17032.
[12]. Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. Journal of machine learning research, 22(268), 1-8.
[13]. Sutton, R. S. (1995). Generalization in reinforcement learning: Successful examples using sparse coarse coding. Advances in neural information processing systems, 8.

