Warsaw University of Technology
Faculty of Electrical Engineering

Comparing the Efficiency of Selected Reinforcement learning Algorithms in Stability Control and Navigation Tasks

Authors: Oskar Wyłucki BEng, Radosław Roszczyk PhD

Reinforcement learning (RL) enables agents to learn through interaction with their environment by maximizing cumulative rewards. This research compares the performance of four popular RL algorithms—DQN, A2C, REINFORCE, and PPO—across two environments of varying complexity: Lunar Lander and Cart Pole. The goal is to evaluate algorithm efficiency, stability, and adaptability.

Environments

Lunar Lander

Lunar Lander is a complex task requiring a lander to safely touch down using main and side thrusters. The agent observes position, speed, angle, and leg contact. Rewards depend on landing precision and fuel use, with penalties for crashes. This environment is non-linear and stochastic, making it significantly more challenging.

Cart Pole

Cart Pole is a simple control task where an agent balances a pole on a moving cart by moving it left or right. The state includes position, velocity, and pole angle. The task is solved if the pole stays upright for 500 steps. It is a deterministic and low-complexity environment, useful for benchmarking.

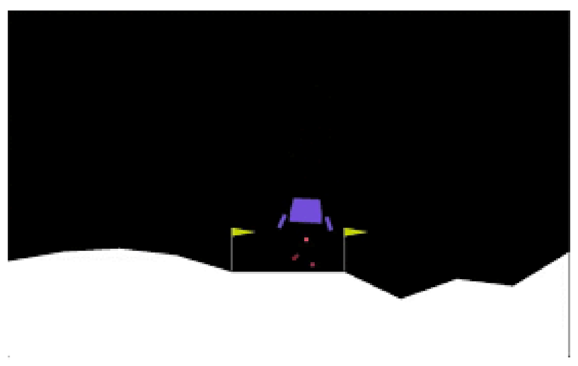

Figure 1. Lunar Lander

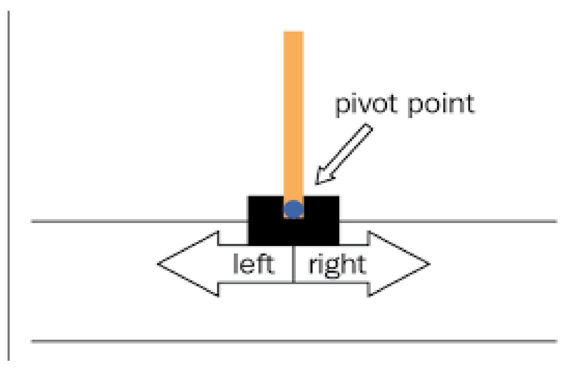

Figure 2. Cart Pole

Algorithms

Deep Q-Learning approximates the action-value function Q(s,a) using a neural network. It features epsilon-greedy exploration, experience replay buffer for random sampling, and a target network for stability. The algorithm iteratively updates network parameters by minimizing the difference between current Q-values and target values calculated using the Bellman equation.

A2C combines two networks: an actor that selects actions according to a learned policy, and a critic that evaluates states. It uses the advantage function $A(s,a) = Q(s,a) - V(s)$ to reduce variance in policy updates. The actor is updated to increase the probability of actions with higher advantages, while the critic is trained to better estimate state values.

REINFORCE is a policy gradient method that directly optimizes policy parameters without learning value functions. It collects complete episodes, calculates cumulative discounted rewards, and updates the policy to increase the likelihood of actions that led to higher returns. However, it suffers from high variance as it lacks a baseline for comparison.

PPO improves training stability by limiting the size of policy updates through a clipped surrogate objective function. It reuses collected data multiple times while preventing excessive policy changes that could destabilize learning. PPO employs Generalized Advantage Estimation for better reward attribution and adds an entropy term to encourage exploration.

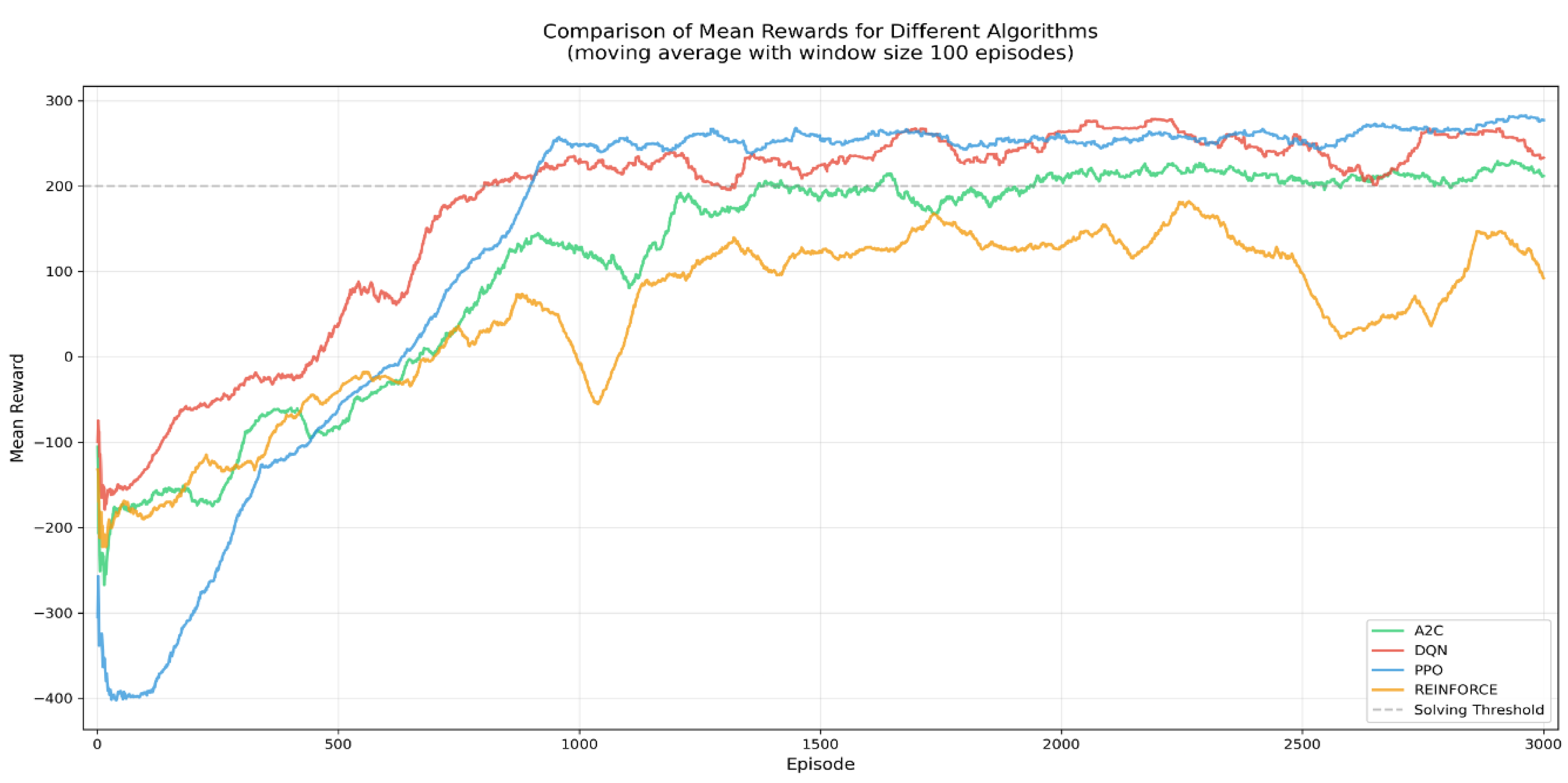

Figure 3. Lunar Lander training results

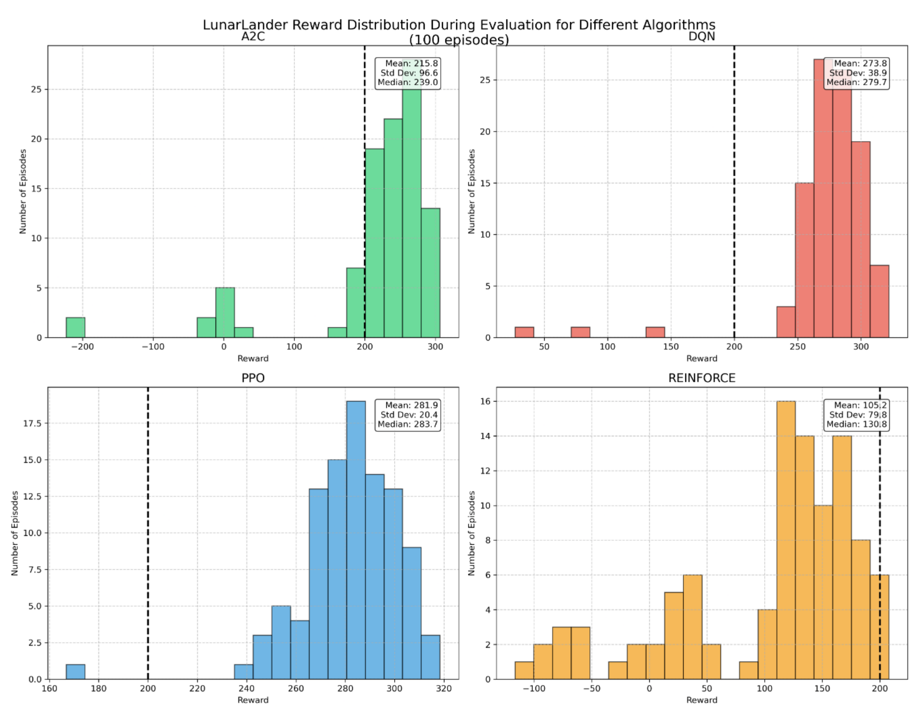

Figure 5. Lunar Lander test results

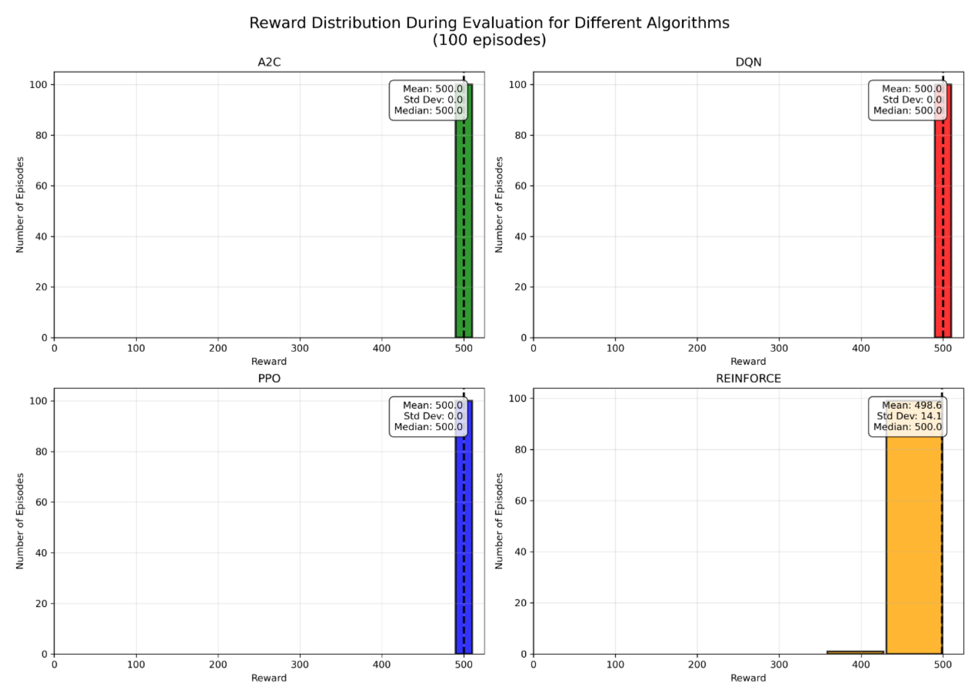

Figure 6. Cart Pole test results

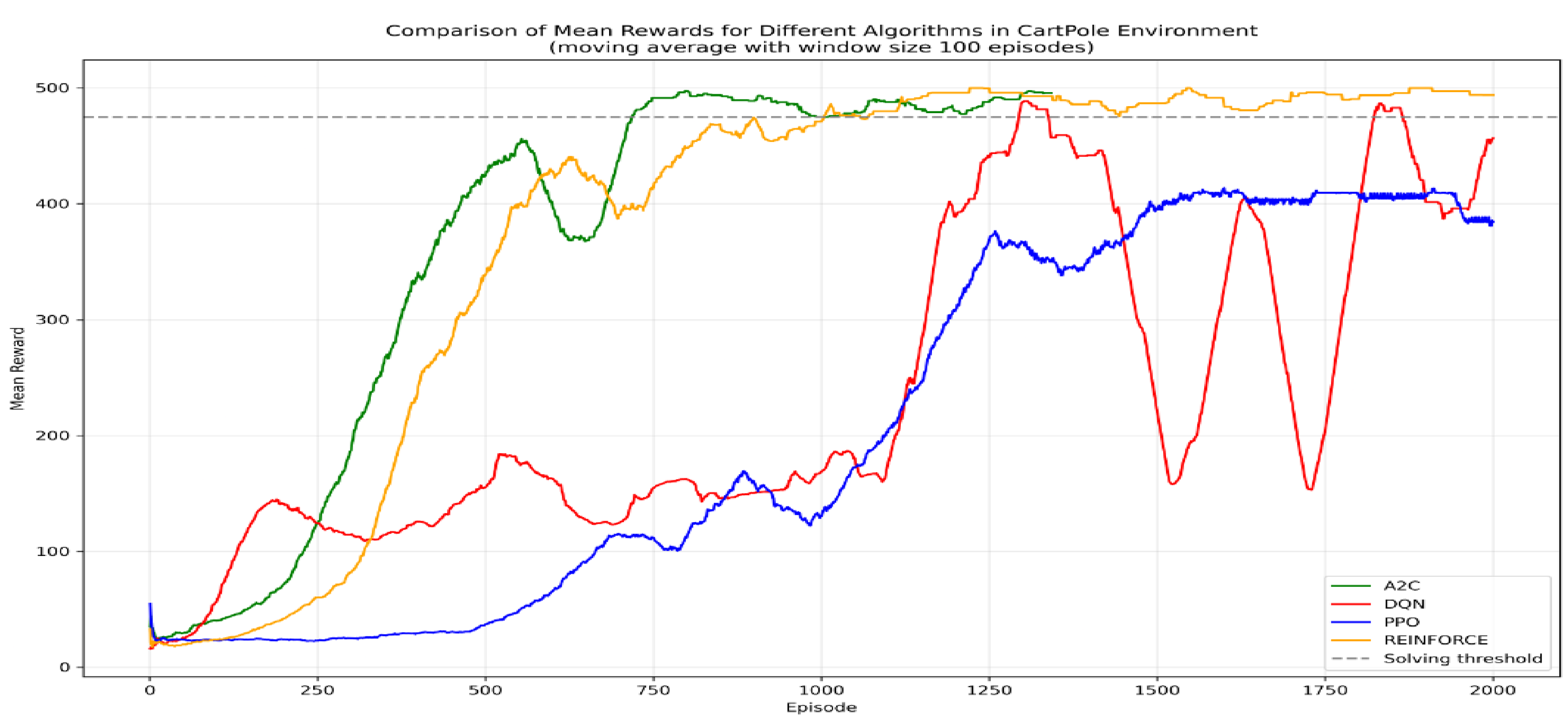

Figure 4. Cart Pole training results

Metric	Algorithm			
	PPO	DQN	A2C	REINFORCE
Lunar Lander				
Mean Reward	281.9	273.8	215.8	105.2
Standard Deviation	20.4	38.9	96.6	79.8
Execution Time	25m 29s	161m 13s	194m 18s	143m 27s
Cart Pole				
Mean Reward	500.0	500.0	500.0	490.0
Standard Deviation	0.0	0.0	0.0	10.5
Execution Time	5m 48s	70m 1s	132m 19s	96m 27s
Key Characteristics				
Learning Stability	High	Medium	Medium	Low
Complex Environment Efficiency	High	Medium	Medium	Low
Experience Utilization	Multiple	Multiple	Single	Single
Implementation Complexity	High	Medium	Medium	Low

Figure 7. Performance comparison

Conclusions and observations

In the Cart Pole environment, all four algorithms successfully solved the task. A2C demonstrated the fastest learning, achieving high scores early due to efficient variance reduction. REINFORCE, despite a slower start, achieved the most stable results by the end of training. DQN exhibited irregular fluctuations, possibly due to overfitting or instability in the Q-network. PPO showed steady convergence, though slower due to its more complex update mechanisms. These differences highlight how algorithm characteristics influence learning dynamics, raising the question of whether to prioritize rapid progress or long-term stability in simpler control tasks. In the more demanding Lunar Lander environment, PPO outperformed other algorithms with the highest mean reward and lowest standard deviation, indicating both effectiveness and training stability. DQN initially performed well but suffered from instability later. A2C achieved consistent results but lagged in peak performance. REINFORCE struggled due to high gradient variance and lack of baseline functions. These performance gaps illustrate how environmental complexity magnifies the strengths and weaknesses of different approaches. As task complexity increases, variance reduction techniques and controlled policy updates. become more critical, potentially outweighing considerations of algorithmic simplicity.