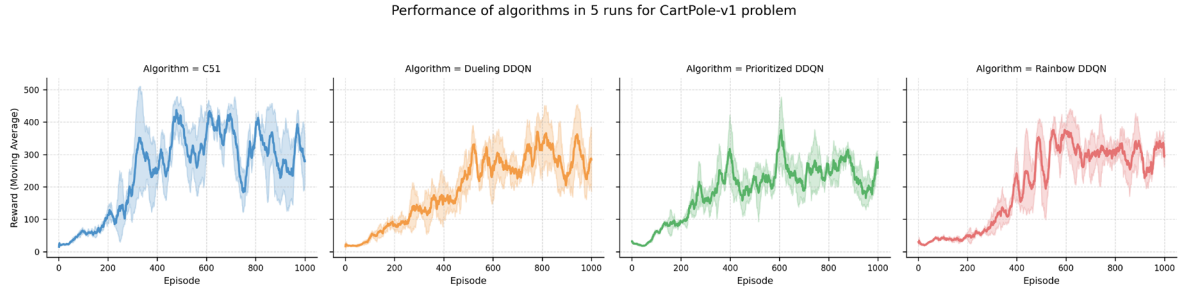
Table S1 Approximate number of training episodes required for each technique to reach the convergence threshold in each environment, based on the moving average reward (rolling window = 20)

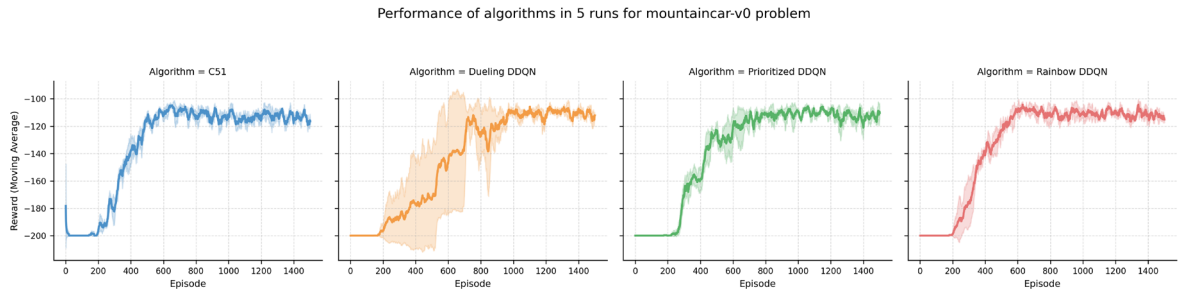
Technique	Lunar Lander (90% of the optimal reward= 180)	Cart Pole (90% of the optimal reward= 450)	Mountain Car (90% of the optimal reward= -120)
DDQN	~650	Not Reached	Not Reached
1 AdvB	~450	Not Reached	~1350
5 AdvB	~450	~680	~1250
10 AdvB	~400	~530	~1150

Episode values were determined from the first point at which the moving average reward (rolling window = 20) reached the predefined convergence threshold. For LunarLander-v2 and CartPole-v1, the convergence thresholds were defined as 90% of the optimal rewards of 200 and 500, corresponding to thresholds of 180 and 450, respectively. For MountainCar-v0, due to the negative reward structure, the convergence threshold was defined as 90% progress toward the solved threshold, resulting in a threshold of -120. "Not Reached" indicates that the threshold was not achieved within the training horizon

Supplementary Figure S1. Comparison of C51, Dueling DDQN, Prioritized DDQN, and Rainbow DDQN in the CartPole-v1 environment.



Supplementary Figure S2. Comparison of C51, Dueling DDQN, Prioritized DDQN, and Rainbow DDQN in the MountainCar-v0 environment.



Supplementary Figure S3. Comparison of C51, Dueling DDQN, Prioritized DDQN, and Rainbow DDQN in the LunarLander-v2 environment.

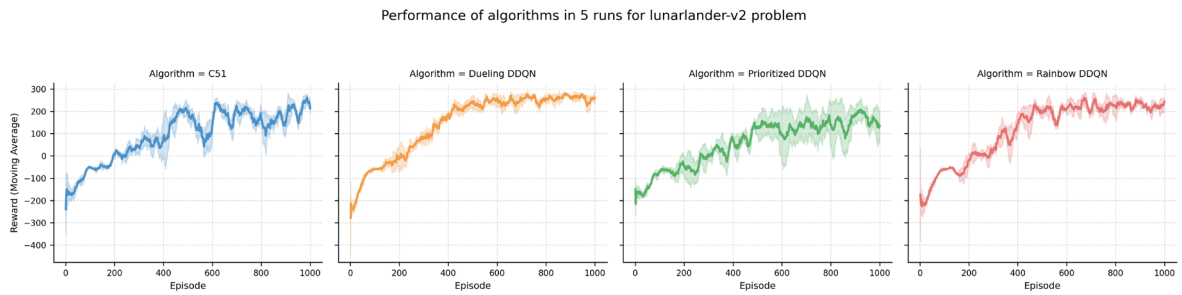


Figure S4. Effect of different advisory-board update cycles on AdvB-RL performance.

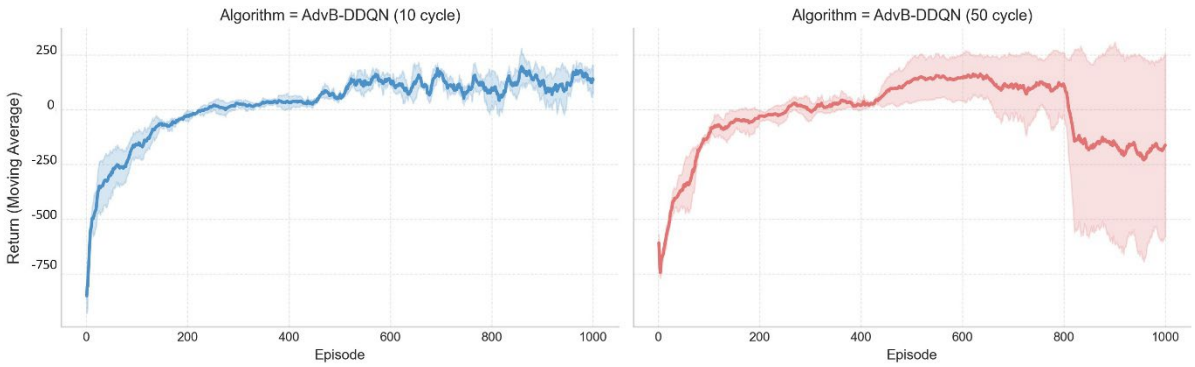


Figure S5. Effect of different cooperative action-selection policies on AdvB-RL performance.

