# 合作傾向建模以緩解多智能體強化學習中的停滯問題

# *Modeling Cooperative Tendencies to Mitigate Stagnation in Multi-Agent Reinforcement Learning*

指導教授: 蔣惟丞副教授　　組別: 計算機工程組　　學生: 黃至鵬、王功羨

## Abstract

In cooperative Multi-Agent Reinforcement Learning (MARL) environments without communication, the absence of intent exchange often causes coordination failures and unstable training. To address this issue, we propose a deep MARL framework that combines a Cooperation Tendency Network (CTN) with Double Deep Q-Networks (DDQN). The CTN estimates the cooperation tendency of state–action pairs, guiding agents away from actions likely to cause coordination breakdowns. To integrate CTN and DDQN, we design an Action Decision Function (ADF) that allocates their contributions. This mechanism emphasizes cooperation-driven exploration in the early training phase and value-based policy learning in later stages. Theoretical analysis shows that the framework stabilizes training and reduces the risk of coordination traps. Experimental results further demonstrate improved sample efficiency and superior final performance compared with conventional approaches in cooperative MARL tasks without communication.

## Architecture

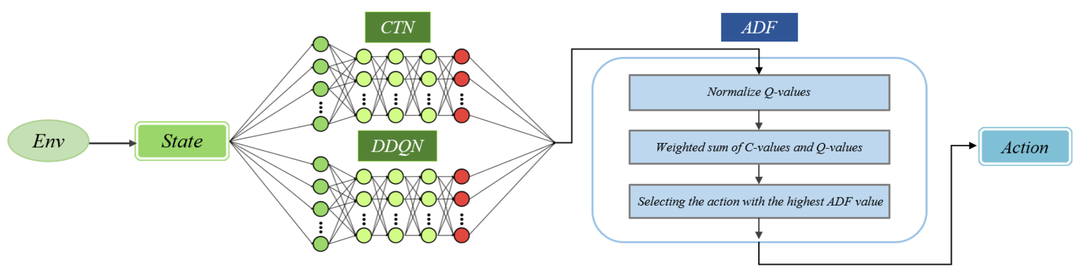


**Fig1. The architecture of Dual Network integrating DDQN and CTN**

## Cooperation Tendency Network (CTN)

In most cooperative scenarios, conventional reinforcement learning (RL) algorithms inevitably encounter stagnation. During the early stages of the training process, agents lack a shared consensus for cooperation, often resulting in decision-making behaviors that are detrimental to collective interests. Our objective is to ensure that agents consistently exploit decision policies that successfully promote cooperation.

To enhance coordinated behaviors among agents, we introduce the Cooperation Tendency Network (CTN). This multilayer perceptron maps each agent's state-action pair to a real-valued scalar between zero and one, formally defined as:

$$C_i\colon \boldsymbol{S_i} \times \boldsymbol{A_i} \to [\mathbf{0}, \mathbf{1}] \in \mathbb{R}$$

where $C_i$ represents the **cooperative tendency value**. A higher cooperative tendency value indicates a greater likelihood that the selected action will support cooperation and prevent coordination failure.

At each policy exploitation step, agents first input the transition tuple $(s_i, a_i, s'_i, r)$ into both networks—the Q-value network and the CTN. Subsequently, the Action Decision Function (ADF) is employed to combine the outputs: the Q-value and the C-value. The agent then selects the action corresponding to the highest combined term to explore the environment.

## Learning Process

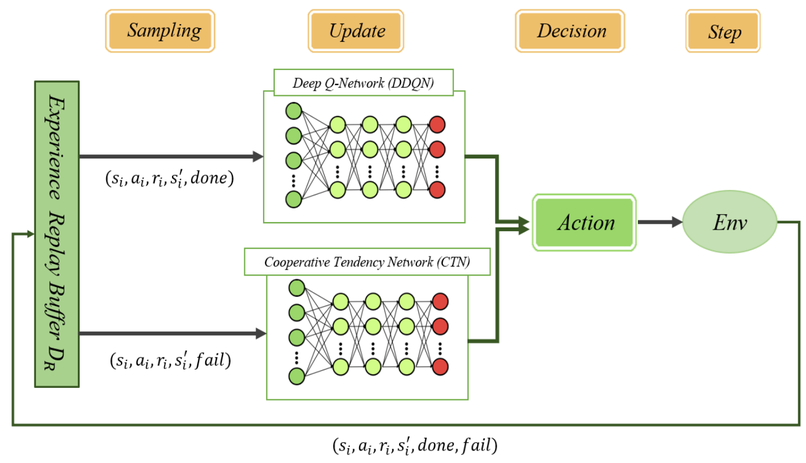


**Fig.2. Training Process of CTN-Q**

## Experiment Setting

The environment is defined as a 20×20 discrete grid. Two cooperative predators start at adjacent positions in the lower-left corner, while the prey is located near the upper-right region. The predators aim to jointly capture the prey, which is achieved when both agents simultaneously occupy positions whose Manhattan distance from the prey is less than or equal to one.

Each predator has a discrete action space consisting of four possible moves: {up, down, left, right}. The environment features two types of prey behavior:

- **Fixed trajectory** – the prey moves cyclically along a square path.
- **Random strategy** – the prey moves or stays in place according to a biased random probability.

The two predators must avoid colliding with each other or moving outside the grid boundaries. If an illegal move occurs, the agent returns to its previous position and receives a penalty.

The reward function is defined as follows:
A small step penalty of −0.1 for each move.
A large reward of +50 upon successfully capturing the prey.
An additional penalty of −10 for collisions or illegal cooperative behavior.

The episode terminates when the prey is captured or when the maximum step limit is reached.

## Results- Fixed trajectory

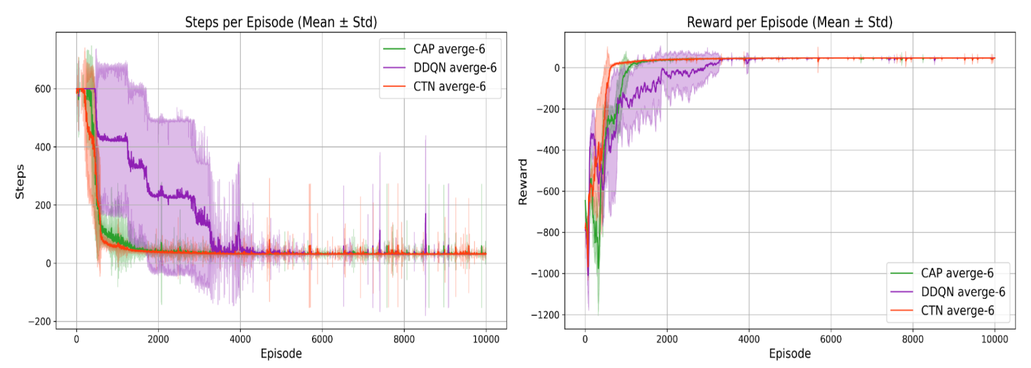


**Fig3. Training Results of Prey-Predators-Fixed trajectory Environment**

## Results- Random strategy

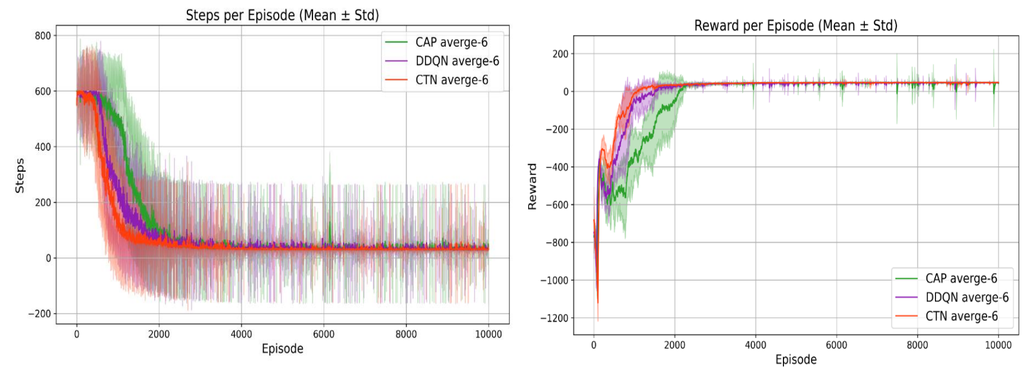


**Fig4. Training Results of Prey-Predators-Random strategy Environment**

## References

[1] Xu, M., Chen, X., She, Y., Jin, Y., Zhao, G., & Wang, J. (2024). *Strengthening cooperative consensus in multi-robot confrontation. ACM Transactions on Intelligent Systems and Technology.*

[2] H. Shi, L. Zhai, H. Wu, M. Hwang, K. S. Hwang and H. P. Hsu (2020), A Multitier Reinforcement Learning Model for a Cooperative Multiagent System, *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 3, pp. 636-644, 2020.

[3] Albrecht, S. V., Christianos, F., & Schäfer, L. (2024). *Multi-agent reinforcement learning: Foundations and modern approaches*. MIT Press.

[4] Bowling, M., & Veloso, M.(2002), Multiagent Learning Using a Variable Learning Rate, *Artificial Intelligence*, vol. 136, no. 2, pp. 215-250, 2002.

[5] Busoniu, L., Babuska, R., & De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 38(2)*, 156–172.

[6] Qiao, H., Rozenblit, J., Szidarovszky, F., & Yang, L. (2006), Multi-agent learning model with bargaining, *Proceedings of the Winter Simulation Conference*, 934–940.