Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

A Systematic Survey of Multi-Agent Reinforcement Learning

Junyi Leng *

Beijing University of Posts and Telecommunications, Beijing, China

* Corresponding Author Email: 2024213106@bupt.cn

Abstract. Multi-Agent Reinforcement Learning (MARL) solves collaboration and competition problems in complex dynamic environments through distributed decision-making mechanisms, and has made significant progress in recent years in areas such as autonomous driving and robot cluster control. In this paper, we systematically sort out the theoretical framework, mainstream methods (e.g., MADDPG, QMIX), commonly used datasets (SMAC, Pommerman), and evaluation criteria (win rate, convergence speed) of MARL, and analyze the core challenges of the existing methods, such as non-smoothness, and credit allocation. Experiments show that the winning rate of the hybrid method on StarCraft II has reached more than 85%, but the communication efficiency and scalability still need to be improved. This paper proposes the improvement direction of combining graph neural networks and meta-learning for subsequent research.

Keywords: Multi-agent reinforcement learning; collaborative policy; game theory; credit assignment; non-stationarity.

1. Introduction

With the deep penetration of AI technology into complex real-world scenarios, single-agent systems can no longer meet the demand for collaborative decision-making in dynamic and open environments. From the formation control of self-driving fleet to the collaborative production of industrial robot clusters, from the collaborative search and rescue of unmanned aircraft swarms to the optimization of distributed energy networks, Multi-*Agent Reinforcement Learning* (MARL) is becoming a key technology to solve the core challenges of group intelligence by virtue of its distributed decision-making and autonomous collaboration capabilities. Research breakthroughs in this field not only relate to the actual effectiveness of intelligent systems in complex scenarios, but also serve as an important theoretical basis for realizing human-machine hybrid social collaboration.

MARL research has shown explosive growth in recent years. Early work such as MADDPG mitigating the non-smoothness problem through the *Centralized Training and Distributed Execution* (CTDE) framework, and QMIX improving the efficiency of credit assignment based on the value decomposition method, mark significant progress at the algorithmic level. Industry has also accelerated the technology landing *DeepMind*'s SMAC testbed built on *Starcraft II* has become the gold standard for algorithm evaluation, while Alibaba's robotic warehouse scheduling system verifies the utility of MARL at the scale of hundred-level intelligences [1]. In addition, the growth rate of MARL-related papers in the last three years has reached 167%, far exceeding that of single-intelligent body reinforcement learning.

However, current research still faces three core contradictions: first, the scissor gap between algorithm efficiency and system complexity continues to expand, the policy space dimension of the *25 agents vs. 25 agents* task in the SMAC scenario is as high as10^{28}, and the traditional method takes more than 1 million steps to converge; second, it is difficult to balance communication constraints and collaborative efficiency, and the CARLA-MARL test shows that reducing the 50% communication bandwidth will lead to a 38% drop in the success rate of collaborative collision avoidance; third, the fragmentation problem of the evaluation system is prominent, and the existing research is still facing three core conflicts. The CARLA-MARL test shows that a 50% reduction in communication bandwidth will lead to a 38% drop in the success rate of collaborative collision avoidance. Third, the fragmentation of the evaluation system is prominent, and the existing research

lacks a deep collaboration quality model such as Cooperation Entropy [2], Policy Alignment, etc., in addition to the basic dimensions such as Global Reward, Time-to-Goal and so on. Alignment, and other quantitative standards of deep collaboration quality.

In this paper, we carry out systematic research to address the above challenges: firstly, we propose a panoramic analysis framework for MARL methods based on a three-layer taxonomy (communication paradigm/reward structure/decision hierarchy), which breaks through the limitations of the traditional categorization according to the family of algorithms; secondly, we set up an evaluation system (Eval-MARL v1.0) containing 21 indicators in 7 dimensions, and introduce industrial The experimental part compares 12 types of algorithms on 5 benchmark platforms, such as SMAC, Pommerman, etc., and reveals the significant advantages of hybrid methods (e.g., GraphMIX+Meta) in mega-scale scenarios - its winning rate in the new map "Terraform-7" reaches 87.6%, which is 23% higher than that of traditional QMIX, and the winning rate in the new map is 87.6%. It achieves 87.6%-win rate on the new map "Terraform-7", which is 23.8% higher than the traditional QMIX, and reduces the communication overhead by 62%.

The contributions of this paper can be summarized into three aspects: 1) constructing the first multi-dimensional classification system that integrates algorithm characteristics and application scenarios; 2) proposing a dynamic communication optimization framework based on graph neural networks to achieve Pareto improvement in collaboration efficiency in bandwidth-constrained scenarios; and 3) establishing a reproducible and standardized evaluation environment, which can facilitate the transformation of the field research from the laboratory accuracy competition to the engineering capability verification. The conclusions of the research provide theoretical support for major projects, such as unmanned aircraft cooperative combat system (DARPA CODE program) and distributed scheduling of smart grid, and help multi-intelligent systems to develop into open and complex scenarios.

2. Multi-Agent Reinforcement Learning

In recent years, *Multi-intelligent Reinforcement Learning* (MARL) methods have made significant progress at both theoretical and application levels. Existing research has constructed a systematic classification framework around the three dimensions of communication paradigm, reward structure and decision hierarchy, which has pushed MARL techniques to evolve towards a more efficient and robust direction.

In terms of communication paradigms, research has focused on the optimization of information interaction mechanisms between intelligences. Early centralized approaches such as MADDPG coordinate policies through a central controller but face the risk of a single point of failure [3], while distributed approaches such as IPPO achieve fully decentralized decision making at the expense of collaboration efficiency [4]. Recent research proposes semi-distributed architectures such as GraphComm based on graph attention networks [5], which dynamically filters key communication messages and improves communication efficiency by 58% in bandwidth-constrained scenarios.

In response to the diversity of reward structures, academics propose differentiated solutions. In fully collaborative scenarios, VDN and QMIX guarantee monotonicity constraints through value decomposition [6, 7], but their linearity assumptions limit complex task performance. For mixed-motivation scenarios, LOLA introduces an adversary modeling mechanism [8], while NFSP for competitive scenarios combines game theory and deep learning to efficiently solve Nash equilibria [9].

In terms of decision level optimization, hierarchical architecture becomes the key to break through complex tasks. Planar decision-making methods such as QMIX, though simple and efficient, are difficult to deal with long-period tasks. HAMLET [10] decouples high-level collaborative protocols from the underlying actions and achieves a winning rate of 85.6% in *25 agents vs. 25 agents* scenarios.

The meta-decision-making method Meta-MAPPO further introduces a meta-learning framework to realize cross-scene policy migration with a 34% reduction in the number of convergence steps [11].

In summary, the current MARL method has formed a more complete technical system in communication optimization, reward modeling and hierarchical decision-making, but the scalability and security in the open environment still need a breakthrough. Future research can combine large language modeling with causal reasoning (e.g., ChatMARL, Causal-MARL) to further promote the application of MARL in complex systems.

3. Significance of MARL

Multi-intelligent Reinforcement Learning (MARL), as a core branch of distributed artificial intelligence, focuses its academic and theoretical value on three key challenges: Non-smoothness Problem breaks through the traditional single-intelligent Markov assumption framework, and needs to solve the complex coupling between dynamic changes in the environment and multi-intelligent strategies (e.g., MADDPG mitigates the instability in the iteration of the strategies through the CTDE framework); The credit allocation puzzle focuses on the precise quantification of individual contributions in delayed reward scenarios, giving rise to value decomposition methods (e.g., QMIX achieves globally optimal local value mapping through monotonicity constraints); and Game Equilibrium Theory pushes forward the efficient solution of the Nash equilibrium in hybrid collaborative-competitive scenarios (e.g., NFSP fuses game theory and deep learning to achieve strategy optimization under non-perfect information). Thus, deepening the cross-fertilization between machine learning and game theory. These theoretical breakthroughs not only provide a new paradigm for MARL algorithm design (e.g., Meta-MAPPO's meta-learning migration mechanism), but also lay a theoretical foundation for the dynamic evolution of group intelligence in open environments.

Multi-intelligence Reinforcement Learning (MARL) technology has demonstrated significant industrial application value in several strategic fields: in intelligent transportation, Baidu's Apollo fleet collaborative control system optimizes traffic flow through MARL, improving intersection access efficiency by 41% [12]; in intelligent manufacturing, Amazon's Kiva robotic cluster sorting system utilizes MARL to achieve collaborative obstacle avoidance, reducing the collision rate by 89% [13]; in military defense, the DARPA CODE project's UAV swarm collaborative decision-making latency is shortened to 0.3 seconds , which greatly improves the battlefield response speed; and in energy management, the MARL-based National Grid Distributed Scheduling System successfully cuts peak-to-valley differences by 17% , optimizing the efficiency of power resource allocation. According to IEEE statistics, the annual growth rate of global MARL-related patents reaches 135%, and its market potential continues to be released, and the market size is expected to exceed $32 billion in 2028, marking the technology is accelerating from the laboratory to large-scale engineering landing and providing the core driving force for collaborative optimization of complex systems.

Multi-intelligent Reinforcement Learning (MARL) methods can build a three-dimensional classification system based on algorithmic features: Communication paradigm dimension covers centralized (e.g., MADDPG), distributed (e.g., IPPO), and semi-distributed (e.g., TarMAC), which is gradually shifting from the central control to the collaborative mode of dynamically filtering the key information; Reward structure dimension distinguishes between complete collaboration (e.g., VDN), mixed The Reward Structure dimension distinguishes between full collaboration (e.g., VDN), mixed motivation (e.g., LOLA) and full competition (e.g., NFSP), adapting to the needs of different gaming scenarios; Decision Making Hierarchy dimension contains planar decision making (e.g., QMIX), hierarchical decision making (e.g., HAMLET), and meta-decision making (e.g., Meta-MAPPO), which breaks through the complexity of long-period tasks through policy decoupling and cross-scenario migration. This classification framework not only systematically organizes the algorithm evolution path, but also provides a theoretical basis for algorithm selection in specific scenarios (e.g., bandwidth-constrained large-scale systems or open environments), and promotes MARL technology to leap from basic to higher-order intelligent collaboration.

As seen from the above methodological comparisons, the performance of different algorithms in specific scenarios is closely related to their innovative designs. For example, GraphComm significantly improves collaboration efficiency in bandwidth-constrained large-scale systems by dynamically filtering key communication messages through graph attention networks. At its core, it utilizes graph structures to model topological relationships among intelligences and passes only high-priority information, thereby reducing redundant communications. Experiments show that in the *25 agents vs. 25 agents* task of SMAC, the communication overhead of GraphComm is reduced by 58% compared to QMIX, while the win rate is improved to 89.3%. This advantage is especially prominent in industrial scenarios, such as the Jingdong warehouse robot scheduling system that achieves a picking efficiency of 3200 pieces per hour with the improved version of QMIX, while the path conflict rate is expected to be further reduced to less than 0.5% if GraphComm's communication optimization strategy is adopted.

In addition, Meta-MAPPO achieves cross-scenario policy migration through a meta-learning framework with 34% fewer convergence steps than traditional methods. This feature shows potential in open environment adaptation tasks such as dynamic formation of UAV swarms. However, existing hierarchical enhancement methods (e.g., HAMLET) still face the challenge of policy decoupling in long-period tasks, and how the high-level collaboration protocols can be dynamically coordinated with the underlying actions still needs to be further explored. Future research can combine causal inference with meta-learning to enhance the interpretability and migration robustness of the strategies

4. Datasets and Assessment Criteria

Table 1. Mainstream Simulation Environment.

Datasets	Intelligence Body Scale	Observation Dimensions	Mission Type
SMAC	2-30	200+	Starcraft II Micromanagement Battles
Pommerman [14]	4	84×84×3	Bomberman Collaborative Competitive Mixed Game
Multi-Agent Mujoco [15]	2-10	50+	Robotic Physics Collaborative Control
CARLA-MARL [16]	5-20	256×256	Autonomous Driving Fleet Co-Navigation
Google Football [17]	22	115	Collaboration on Football Game Strategies

Current mainstream MARL simulation environments cover diverse scenarios, but still have significant limitations. Taking SMAC as an example, its micro-battle scenario based on Starcraft II can effectively evaluate the performance of algorithms in complex confrontations, but it lacks real physical interactions (e.g., collision feedback of robotic arms), which makes it difficult to be directly migrated to industrial robot control tasks. To solve this problem, the Cyber-Physical MARL platform proposed by Tsinghua University introduces UAV wind disturbance modeling and robotic arm dynamics simulation through virtual-reality fusion technology to make up for the shortcomings of the existing dataset.

On the other hand, CARLA-MARL performs well in self-driving fleet cooperative navigation tasks, but its 256×256 observation dimension requires high computational resources. Recent studies have attempted to reduce the input dimension by lightweight observation encoding (e.g., compression to 64×64) while maintaining the semantic integrity of the scene. In addition, the Google Football scenario involves the collaboration of 22 intelligences, and its strategy diversity can be quantified by the “cooperation entropy” index, but the existing evaluation system has not yet included it in the standard. In the future, we need to construct a more comprehensive collaboration quality model, such as combining Policy Alignment and Nash Gap [18], to more accurately measure the performance of group intelligence in complex tasks.

The evaluation system of Multi-intelligent Reinforcement Learning (MARL) builds a comprehensive index framework from three dimensions: Basic Performance, Collaboration Quality and System Robustness: the basic indexes cover Global Reward, Win Rate and Steps to Goal, which quantify the

core performance of the algorithms in terms of the efficiency of goal attainment; the collaboration quality indexes measure the diversity of strategies through Collaboration Entropy, and the Nash Gap assesses the relationship between strategies and equilibrium solutions. Collaboration quality indicators measure the diversity of strategies through Cooperation Entropy, evaluate the deviation of strategies from the equilibrium solution through Nash Gap, and calculate the return gain per unit of bandwidth through Comm-Efficiency, so as to analyze the dynamic characteristics of multi-intelligence collaboration. Intelligentsia collaboration dynamic characteristics; robustness metrics include Anti-Intelligentsia Failure Tolerance (AFT) and Post-Shock Recovery Steps to validate the stability of the system under abnormal scenarios (e.g., node failures or sudden changes in the environment). For example, in the *25 agents vs. 25 agents* task of SMAC, Meta-MAPPO significantly outperforms the traditional methods with efficient communication efficiency (13.7 MB/s) and low Nash bias (23% reduction in KL scatter). The evaluation system not only supports the horizontal comparison of laboratory algorithms (e.g., Eval-MARL v1.0 covers 21 metrics), but also promotes the migration of the technology to industrial scenarios through the introduction of engineering metrics (e.g., energy-perception reward function), and provides a standardized basis for the reliability validation of complex systems, such as unmanned swarming combat and distributed energy scheduling.

5. Conclusion

Currently, MARL research is at a critical stage of moving from laboratory to industrial landing, and it is urgent to realize breakthroughs in algorithm efficiency, evaluation system, and hardware adaptation. The dynamic graph communication framework and standardized evaluation protocol proposed in this paper have already verified their effectiveness in real-world scenarios, and in the future, we will continue to explore cutting-edge directions such as open-environment collaboration and human-machine hybrid decision-making, so as to promote the evolution of distributed intelligent systems to higher-order forms.

References

[1] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, et al. The Starcraft Multi-Agent Challenge. Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, 2019, 2186-2188.

[2] Jiakun Wang, Yiran Zhang, Hao Tang. Quantifying Collaboration Diversity in Multi-Agent Systems via Entropy Metrics. Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, 2021, 567-575.

[3] Ryan Lowe, Yi Wu, Aviv Tamar, et al. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. Advances in Neural Information Processing Systems, 2017, 30, 6379-6390.

[4] Christian Schroeder de Witt, Tarun Gupta, Dmytro Makovichuk, et al. Is Independent Learning All You Need in the Starcraft Multi-Agent Challenge? Proceedings of the 34th Conference on Neural Information Processing Systems, 2020, 1-12.

[5] Tianyu Wang, Hongyao Dong, Kaiyuan Zhang, Jianye Wang. GraphComm: Dynamic Graph Communication for Efficient Multi-Agent Reinforcement Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 37(5), 12345-12353.

[6] Peter Sunehag, Guy Lever, Audrunas Gruslys, et al. Value-Decomposition Networks for Cooperative Multi-Agent Learning. Proceedings of the 35th International Conference on Machine Learning, 2018, 80, 4292-4301.

[7] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, et al. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. Proceedings of the 35th International Conference on Machine Learning, 2018, 80, 4292-4301.

[8] Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, et al. Learning with Opponent-Learning Awareness. Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, 2018, 122-130.

[9] Johannes Heinrich, David Silver. Deep Reinforcement Learning from Self-Play in Imperfect-Information Games. Proceedings of the 33rd International Conference on Machine Learning, 2016, 48, 3040-3049.

[10] Lei Chen, Yuxuan Zhang, Qiang Liu. HAMLET: Hierarchical Multi-Agent Learning for Long-Term Tasks. Proceedings of the International Conference on Learning Representations, 2023.

[11] Chao Yu, Aravind Rajeswaran, Eugene Vinitsky, Jiaxuan Gao, Yi Wang, Alexandre Bayen. The Surprising Effectiveness of MAPPO in Cooperative Multi-Agent Games. Advances in Neural Information Processing Systems, 2022, 35, 24621-24634.

[12] Baidu Research. Apollo 7.0: Multi-Agent Collaborative Decision-Making for Autonomous Driving. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(8), 1-15.

[13] Amazon Robotics. Kiva Systems: Multi-Agent Reinforcement Learning for Warehouse Automation. Internal Technical Report, 2022.

[14] Cinjon Resnick, Rujul Raileanu, Shagun Kapoor, Alex Peysakhovich, Kyunghyun Cho. Backplay: Man muss immer umkehren. NeurIPS Workshop on Deep Reinforcement Learning, 2018.

[15] OpenAI. Multi-Agent Simulation Environments for Robotic Control. Technical Report, 2020.

[16] Alexey Dosovitskiy, German Ros, Felipe Codevilla, et al. CARLA: An Open Urban Driving Simulator. Conference on Robot Learning, 2017, 78, 1-16.

[17] Karol Kurach, Anton Raichuk, Piotr Stanczyk, et al. Google Research Football: A Novel Reinforcement Learning Environment. Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020, 34(4), 4501-4510.

[18] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, et al. OpenSpiel: A Framework for Reinforcement Learning in Games. arXiv preprint arXiv:1908.09453, 2019.