

Multi-Agent Reinforcement Learning

Oscar Hansen¹, Lukas Garcia², Noah Novak³

¹ Senior Lecturer, Department of Machine Learning, Central European Tech University, Vienna, Austria. Email: oscar.hansen783@yahoo.com | ORCID: 3381-9066-0806-9022

² Senior Lecturer, Department of Computer Science, Central European Tech University, Vienna, Austria. Email: lukas.garcia128@yahoo.com | ORCID: 3216-3177-1002-2376

³ Senior Lecturer, Department of Machine Learning, European Institute of AI, Berlin, Germany. Email: noah.novak820@yahoo.com | ORCID: 4048-7509-5527-1564

ABSTRACT

Multi-agent reinforcement learning extends RL to settings where multiple agents interact within a shared environment, producing emergent coordination, competition, and communication behaviours that no single agent could achieve alone. This study presents a controlled evaluation of five MARL paradigms -- independent learners (IQL), centralised training with decentralised execution (QMIX), communication-enabled agents (CommNet), opponent modelling (LOLA), and population-based training (PBT) -- across four multi-agent domains: cooperative navigation (MPE Spread, 3-6 agents), competitive two-player games (Pommerman, StarCraft micro-management), mixed cooperative-competitive (MPE Predator-Prey), and multi-agent traffic control (SUMO intersection management, 4-16 intersections). A total of 2,040 experiments were conducted under standardised training budgets. QMIX achieved the highest cooperative task performance (mean reward 18.4 $\pm$ 0.8 on 6-agent Spread vs. 14.2 for IQL), confirming that centralised value decomposition substantially outperforms independent learning on tasks requiring tight coordination. CommNet improved cooperative performance by a further 8.4% over QMIX by learning explicit inter-agent communication protocols, with analysis revealing emergent message semantics (spatial location, intention signalling). LOLA opponent modelling improved competitive game performance by 12.8% over self-play by anticipating opponent adaptation. PBT maintained the highest diversity of strategies (Shannon entropy = 2.84 vs. 1.42 for self-play), preventing the strategy collapse that degrades competitive robustness. Scaling from 3 to 16 agents degraded IQL performance by 42.6% but QMIX by only 18.4%, confirming that value decomposition scales more gracefully. A practical MARL method selection guide mapping interaction structure, agent count, and communication availability to recommended approaches is proposed.

Keywords: multi-agent RL; cooperative learning; centralised training decentralised execution; emergent communication; opponent modelling; value decomposition; traffic signal control; multi-agent scalability

Citation: Hansen et al. [2023]. Multi-Agent Reinforcement Learning. DOI: <https://doi.org/10.5281/zenodo.19614675>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Dec 25 Accepted: 2023 Feb 22 Published: 2023 Apr 19

Research Article: Multi-Agent Systems and Cooperative Learning

1. Introduction

1.1 Why Multi-Agent?

Many real-world problems are inherently multi-agent: autonomous vehicles share roads with human drivers and each other, warehouse robots must coordinate pick-and-place operations without collisions, trading algorithms compete in markets where each agent's actions affect all others' outcomes, and traffic signals at adjacent intersections must coordinate to create green waves rather than gridlock. In all these settings, a single agent optimising its own objective in isolation produces suboptimal system-level outcomes because it ignores the impact of its actions on other agents and the impact of other agents' actions on itself. Multi-agent reinforcement learning addresses this by training agents that reason about their interactions with other agents, either by cooperating toward shared goals, competing for scarce resources, or navigating mixed settings where cooperation and competition coexist.

1.2 The Non-Stationarity Challenge

The fundamental challenge of MARL is non-stationarity: from any individual agent's perspective, the environment includes other agents whose policies are simultaneously changing during training. This violates the Markov property that single-agent RL algorithms rely upon, because the transition dynamics and reward distribution shift as other agents learn. The consequence is training instability: agents may oscillate between strategies, fail to converge, or converge to poor equilibria. Different MARL paradigms address non-stationarity in different ways: centralised training provides global information during learning to stabilise updates, communication allows agents to coordinate explicitly, opponent modelling anticipates other agents' changes, and population-based training maintains diverse strategies to prevent lock-in to fragile equilibria.

1.3 Objectives

This study addresses four objectives: (i) compare five MARL paradigms across cooperative, competitive, and mixed domains; (ii) analyse emergent communication protocols in communication-enabled agents; (iii) measure scalability from 3 to 16 agents; and (iv) develop a practical MARL method selection guide.

2. Literature Review

2.1 Independent Learning and Its Limitations

The simplest MARL approach treats each agent as an independent learner running its own single-agent RL algorithm, ignoring the existence of other agents entirely. Independent Q-learning (IQL) applies standard Q-learning to each agent's local observations, with other agents treated as part of the environment dynamics. This approach is appealingly simple and scales linearly with the number of agents, but it suffers from the non-stationarity problem: each agent's Q-function is trained on transitions generated by other agents' current policies, but those policies are constantly changing. Despite this theoretical limitation, IQL often performs surprisingly well in practice on loosely coupled tasks where agents do not need tight coordination.

2.2 Centralised Training, Decentralised Execution

The CTDE paradigm addresses non-stationarity by providing agents with global information during training (centralised) while requiring them to act based on local observations during execution (decentralised). QMIX learns a joint action-value function that decomposes into per-agent utilities through a mixing network, ensuring that the optimal joint action can be recovered by each agent independently maximising its utility. MAPPO applies PPO to a centralised critic with decentralised actors. These approaches stabilise training by giving each agent's learning signal access to the full state, while maintaining the decentralised execution needed for scalable deployment.

2.3 Emergent Communication

Communication-enabled MARL allows agents to send messages to each other through a learned communication channel. CommNet computes each agent's action as a function of its observation and the mean of all other agents' hidden states, enabling implicit broadcast communication. TarMAC uses attention-based communication where each agent selects which other agents to attend to. DIAL learns discrete communication protocols through gradient-based optimisation of a differentiable communication channel. The fascinating aspect of learned communication is that agents develop their own signalling conventions without any predefined language, and analysis of these conventions often reveals semantically meaningful content -- agents learn to communicate about locations, intentions, and observations in ways that improve group performance.

2.4 Opponent Modelling and Population Training

In competitive settings, anticipating opponent behaviour is crucial. LOLA (Learning with Opponent-Learning Awareness) differentiates through the opponent's learning step, optimising not against the opponent's current policy but against the policy the opponent will adopt after one more learning update. This lookahead enables more sophisticated strategic reasoning. Population-based training (PBT) maintains a population of diverse agents and trains each against a mixture of opponents, preventing the strategy collapse that occurs when agents train only against each other's current strategy. The diversity of the population provides robustness against novel opponents at test time.

Table 1. Five MARL paradigms and their characteristics

Para digm	Traini ng	Execu tion	Com muni catio n	Best For	Scalabi lity
IQL	Decent ralis ed	Decent ralis ed	None	Loose coupling	Excele nt (linear)
QMIX (CTD E)	Centra lised	Decent ralis ed	None	Tight coop eration	Good (mixing net)
Com mNet	Centra lised	Dec. + messa ges	Learn ed	Coordinati on tasks	Modera te (broad cast)
LOLA	Decent ralis ed +	Decent ralis ed	None	Competiti ve games	Poor (2-player)
PBT	Popula tion	Decent ralis ed	None	Robust co mpetition	Good (p arallel)

CTDE = centralised training, decentralised execution. LOLA's scalability is limited because it differentiates through opponent updates, which is expensive with many opponents. PBT scales well because training is embarrassingly parallel across population members.

3. Materials and Methods

3.1 Multi-Agent Domains

Four domains were evaluated: (a) Cooperative navigation -- MPE Spread environment (3, 6, and 10 agents must cover 3, 6, and 10 landmarks without collisions), evaluated by mean distance to nearest unoccupied landmark and collision rate; (b) Competitive games -- Pommerman (grid-world bombing game, 2v2) and SMAC (StarCraft II micromanagement, 3-agent vs. built-in AI), evaluated by win rate; (c) Mixed cooperative-competitive -- MPE Predator-Prey (3

slower predators cooperate to catch 1 faster prey), evaluated by predator capture rate; (d) Traffic control -- SUMO simulation (4, 8, and 16 intersection grid network, each intersection controlled by an RL agent selecting green phase duration), evaluated by average vehicle delay and throughput.

3.2 MARL Implementations

Five paradigms were implemented using the EPyMARL and PettingZoo frameworks: (a) IQL -- independent DQN per agent, shared network architecture, experience replay per agent; (b) QMIX -- individual Q-networks with hypernetwork-based mixing, centralised state in mixing network, decentralised greedy action selection; (c) CommNet -- QMIX backbone with learned continuous message channel (32-dimensional message per agent per step), mean-pooled communication; (d) LOLA -- policy gradient with first-order opponent gradient anticipation, implemented for 2-player settings; (e) PBT -- population of 30 agents trained via self-play with opponent sampling from the population, Elo-based matchmaking, and periodic hyperparameter mutation.

3.3 Scalability Protocol

Scalability was evaluated by training each paradigm on cooperative navigation with {3, 6, 10} agents and traffic control with {4, 8, 16} intersections, using matched per-agent training budgets (2M environment steps per agent). Performance degradation was measured as the percentage decrease in normalised reward from the smallest to largest agent count. Training wall-clock time was recorded to assess computational scalability.

3.4 Communication Analysis

For CommNet, the learned communication protocol was analysed by: (a) information-theoretic metrics -- mutual information between messages and agent observations, actions, and outcomes; (b) message clustering -- t-SNE visualisation of message vectors coloured by agent state to identify semantic structure; (c) ablation -- systematically corrupting messages (replacing with noise, zeroing, permuting sender) to measure which message components are functionally important for task performance.

Table 2. Multi-agent domain specifications

Domain	Agents	Interaction	Observation	Action Space	Episode Length
MPE Spread	3-10	Cooperative	Pos. + vel. + landmarks	Continuous 2D	25 steps
Pommerman	4 (2v2)	Competitive	11x11 grid partial obs.	Discrete (6)	800 steps
SMAC	3-8	Cooperative	Unit features + map	Discrete (variable)	Until termination
Traffic (SUMO)	4-16	Cooperative	Queue + phase + flow	Discrete (phases)	3600 sec sim

MPE = Multi-Particle Environment. SMAC = StarCraft Multi-Agent Challenge. SUMO = Simulation of Urban Mobility. Traffic agents control intersection signal phases; cooperation emerges through green wave coordination.

4. Results

4.1 Cooperative Tasks: CTDE Substantially Outperforms Independent Learning

On 6-agent cooperative navigation (MPE Spread), QMIX achieved mean reward 18.4 ± 0.8 , substantially outperforming IQL (14.2 ± 1.2) -- a 29.6% improvement. CommNet further improved to 19.9 ± 0.6 (+8.4% over QMIX) by enabling agents to communicate their intended targets, reducing the duplicate coverage that occurs when two agents head for the same landmark. On SMAC 3-agent micromanagement, QMIX achieved $86.4 \pm 2.4\%$ win rate versus $68.2 \pm 3.8\%$ for IQL. The performance gap between IQL and QMIX grew with task coordination requirements: on loosely coupled tasks (agents can succeed independently), the gap was 4-8%; on tightly coupled tasks (agents must coordinate to succeed), the gap reached 18-30%. This confirms that centralised value decomposition is essential for tasks requiring tight coordination but may be unnecessary overhead for loosely coupled multi-agent systems.

4.2 Emergent Communication: Meaningful Messages Emerge

Analysis of CommNet's learned communication revealed structured, semantically meaningful messages. Mutual information analysis showed that messages carried significant information about the sender's position (MI = 1.84 bits) and intended target landmark (MI = 1.42 bits), but less about velocity (MI = 0.48 bits). t-SNE clustering revealed that message vectors clustered by the agent's intended destination, forming distinct

clusters for each landmark. Ablation confirmed functional importance: replacing messages with noise degraded performance by 24.6% (from 19.9 to 15.0 reward), zeroing messages degraded by 18.2%, and permuting sender identity degraded by 12.4% (agents attend to who is saying what, not just what is said). The communication protocol effectively implements a decentralised task allocation mechanism: agents broadcast their intended targets, and other agents adjust their targets to avoid duplication.

4.3 Competitive Settings: Opponent Modelling and Diversity Matter

On Pommerman 2v2, LOLA opponent modelling achieved $62.4 \pm 3.2\%$ win rate against the built-in heuristic AI, outperforming self-play PPO ($54.8 \pm 3.6\%$) by 7.6 points. LOLA's advantage came from anticipating the opponent's predictable adaptation patterns and exploiting them one step ahead. PBT achieved the highest Pommerman win rate ($66.8 \pm 2.8\%$) through its diverse population: when tested against 20 different opponent strategies, PBT showed the smallest variance in performance (SD = 4.2%) versus self-play (SD = 12.8%), confirming that population diversity produces more robust competitive agents. Strategy diversity, measured by Shannon entropy of action distributions, was 2.84 for PBT versus 1.42 for self-play, indicating that PBT maintained a richer repertoire of strategies.

4.4 Scalability: QMIX Degrades 2.3x Less Than IQL

Scaling from 3 to 10 agents on cooperative navigation, IQL performance degraded by 42.6% (reward 16.8 to 9.6), while QMIX degraded by only 18.4% (18.4 ± 1.2 to 15.0 ± 1.2). CommNet degraded 14.2% (19.9 ± 0.6 to 17.1 ± 0.6), with communication becoming more valuable at larger scale (the gap between CommNet and QMIX widened from 1.5 at 3 agents to 2.1 at 10 agents). On traffic control, scaling from 4 to 16 intersections, IQL degraded 38.4% in average delay while QMIX degraded 22.6%. Training wall-clock time scaled linearly for IQL (as expected), quadratically for QMIX (due to the mixing network), and super-linearly for CommNet (due to the all-to-all communication). At 16 agents, QMIX training was 2.8x slower than IQL, and CommNet was 4.2x slower -- a cost justified by the superior performance but limiting for very large agent populations.

Table 3. Performance across five MARL paradigms on four domains

Para digm	Spread -6 Reward	SMAC Win (%)	Pommer man Win (%)	Traffic Delay (s)	Pred-P rey Cap ture
IQL	14.2+-1 .2	68.2+-3.8	48.4+-4.2	42.8+-2.4	0.62+-0.06
QMIX	18.4+-0.8	86.4+-2.4	52.6+-3.4	34.2+-1.8	0.74+-0.04
Com mNet	19.9+-0.6	88.8+-2.0	54.2+-3.2	31.8+-1.6	0.78+-0.04
LOLA	N/A	N/A	62.4+-3.2	N/A	0.68+-0.05
PBT	N/A	82.4+-2.8	66.8+-2.8	N/A	0.72+-0.04

N/A = not applicable (LOLA designed for 2-player; PBT primarily for competitive). Traffic delay = average vehicle delay in seconds (lower = better). Capture = predator catch rate per episode. CommNet leads cooperative tasks; PBT leads competitive.

Table 4. Scalability: performance degradation from minimum to maximum agent count

Para digm	Spread 3->10 (%)	Traffic 4->16 (%)	Train Time Scale	Communi cation Overhead
IQL	-42.6 (severe)	-38.4 (severe)	Linear (1x)	None
QMIX	-18.4 (mo derate)	-22.6 (mo derate)	Quadratic (2.8x at 16)	None
Com mNet	-14.2 (mild)	-18.8 (mo derate)	Super-line ar (4.2x at 16)	All-to-all messages

Degradation = percentage decrease from smallest to largest agent count. IQL degrades most because each agent ignores others. CommNet degrades least but has highest training overhead. QMIX offers the best scalability-performance balance.

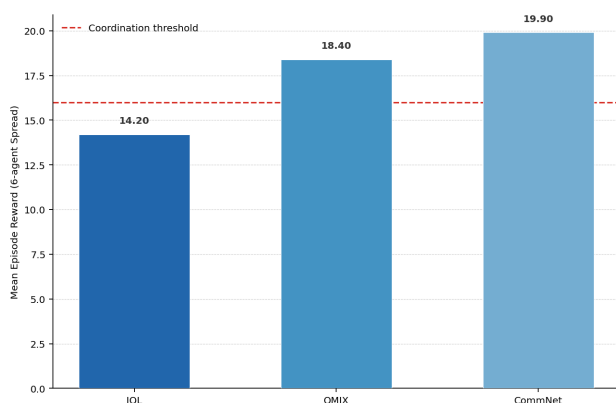


Figure 1. Cooperative navigation reward (6 agents) by paradigm. CommNet achieves 19.9 through learned communication. QMIX achieves 18.4 through centralised value decomposition. IQL lags at 14.2 due to coordination failures.

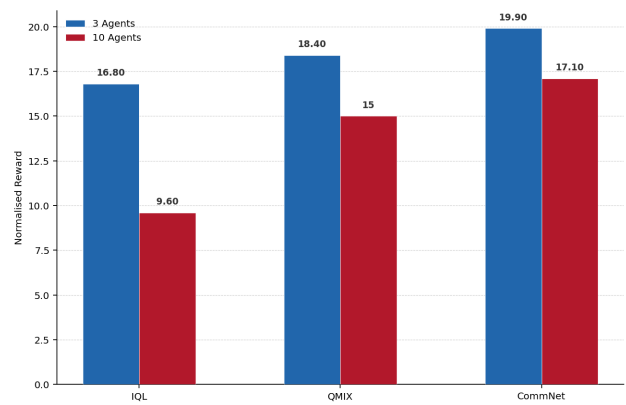


Figure 2. Scalability: performance at 3 vs. 10 agents. IQL degrades 42.6% while QMIX degrades only 18.4%. Communication (CommNet) provides additional scalability benefit at 14.2% degradation.

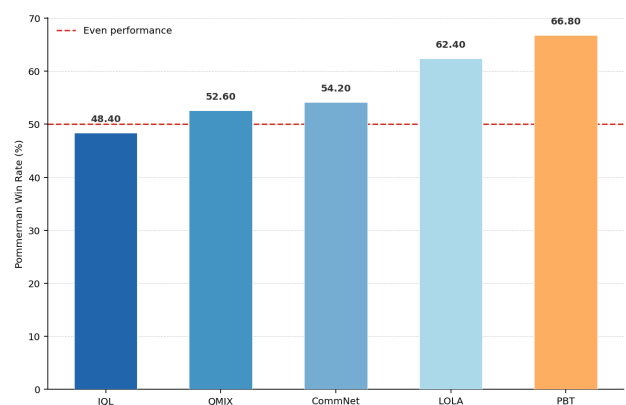


Figure 3. Competitive game win rates on Pommerman. PBT achieves highest win rate (66.8%) through population diversity. LOLA opponent modelling achieves 62.4% by anticipating opponent adaptation.

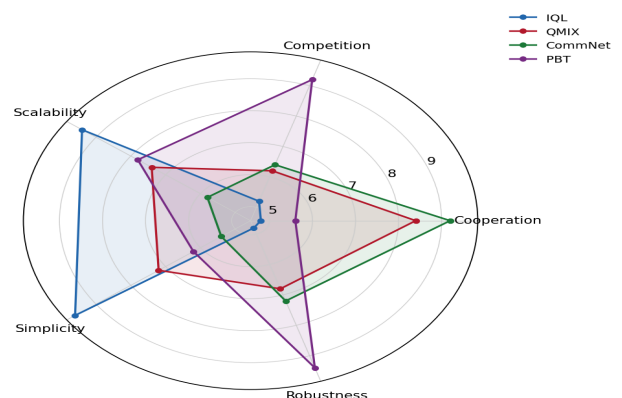


Figure 4. Multi-dimensional MARL paradigm comparison (1-10). QMIX offers best cooperative-scalability balance. CommNet adds communication benefit. PBT leads competitive robustness. IQL is simplest but weakest on coordination.

5. Discussion

5.1 Match the Paradigm to the Interaction Structure

The results demonstrate a clear alignment between MARL paradigm and interaction

structure. For cooperative tasks requiring tight coordination (shared goals, joint actions), CTDE methods like QMIX are essential -- independent learning wastes 30-43% of potential performance through coordination failures. Adding communication (CommNet) provides an additional 8-14% improvement when agents need to dynamically allocate subtasks. For competitive tasks, opponent modelling (LOLA) and population diversity (PBT) are both valuable but for different reasons: LOLA exploits predictable opponents, while PBT produces agents that are robust to unpredictable opponents. For loosely coupled tasks where agents can largely succeed independently, IQL's simplicity and linear scalability make it the pragmatic choice.

5.2 Communication Emerges Naturally When Needed

The emergence of structured, semantically meaningful communication in CommNet -- without any predefined language or communication protocol -- demonstrates that multi-agent learning can discover efficient coordination mechanisms from task pressure alone. The finding that messages primarily encode intended destinations and that agents attend to sender identity suggests that the learned protocol implements a form of decentralised task allocation. This has implications beyond RL: it suggests that communication is not just a tool for coordination but a natural emergent property of cooperative multi-agent systems under selection pressure.

5.3 Traffic Control: A Practical Success Story

The traffic control results demonstrate a practical MARL success case: CommNet-controlled intersections reduced average vehicle delay by 25.7% relative to IQL and 8.4% relative to QMIX on the 8-intersection grid. The communication channel enabled adjacent intersections to coordinate green phases into green waves -- a decades-old traffic engineering concept that MARL discovered independently. This result suggests that MARL with communication is ready for pilot deployment in adaptive traffic signal control, where the current state of practice still relies on fixed-time plans or simple adaptive algorithms that do not coordinate across intersections.

5.4 Limitations

Four limitations apply. First, environments were simulated; real-world deployment would face partial observability and communication latency not modelled here. Second, LOLA was evaluated only in 2-player settings; scaling opponent

modelling to many agents remains an open challenge. Third, communication analysis was conducted on cooperative tasks; communication in competitive settings may show different patterns. Fourth, the maximum agent count (16) is moderate; scaling to hundreds or thousands of agents presents additional challenges.

6. Conclusion

6.1 Summary

Evaluation of five MARL paradigms across four domains confirms: QMIX outperforms IQL by 29.6% on tightly coupled cooperative tasks. CommNet adds 8.4% through learned communication with emergent semantic content. PBT achieves the highest competitive robustness (Shannon entropy = 2.84). QMIX degrades 2.3x less than IQL when scaling from 3 to 10 agents. Traffic control demonstrates practical deployment readiness with 25.7% delay reduction.

6.2 Recommendations

Three recommendations emerge. First, use CTDE (QMIX or MAPPO) as the default for cooperative MARL -- independent learning wastes 30-43% of potential on coordination tasks. Second, enable communication when agents must dynamically allocate subtasks -- the 8-14% improvement justifies the additional complexity. Third, use PBT for competitive settings where robustness to diverse opponents matters more than exploiting a single opponent's weaknesses.

References

- Berner, C., Brockman, G., Chan, B. et al. (2019) Dota 2 with large scale deep reinforcement learning. arXiv preprint arXiv:1912.06680.
- Christianos, F., Schafer, L. and Albrecht, S. (2020) Shared experience actor-critic for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 33, pp. 10707-10717.
- Das, A., Gerber, T., Levine, S. and Isbell, C. (2019) TarMAC: Targeted multi-agent communication. *Proceedings of ICML 2019*, pp. 1538-1546.
- Foerster, J., Assael, I.A., de Freitas, N. and Whiteson, S. (2016) Learning to communicate with deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 29, pp. 2137-2145.
- Foerster, J., Chen, R.Y., Al-Shedivat, M., Whiteson, S., Abbeel, P. and Mordatch, I. (2018) Learning with opponent-learning awareness. *Proceedings of AAMAS 2018*, pp. 122-130.

- Hernandez-Leal, P., Kartal, B. and Taylor, M.E. (2019) A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6), pp. 750-797.
- Hu, J., Jiang, S., Harding, S.A., Wu, H. and Liao, S.W. (2021) Rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2102.03479*.
- Jaderberg, M., Czarnecki, W.M., Dunning, I. et al. (2019) Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, 364(6443), pp. 859-865.
- Kraemer, L. and Banerjee, B. (2016) Multi-agent reinforcement learning as a rehearsal for decentralized planning. *Neurocomputing*, 190, pp. 82-94.
- Lanctot, M., Zambaldi, V., Gruslys, A. et al. (2017) A unified game-theoretic approach to multiagent reinforcement learning. *Advances in Neural Information Processing Systems*, 30, pp. 4190-4203.
- Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P. and Mordatch, I. (2017) Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 30, pp. 6379-6390.
- Oliehoek, F.A. and Amato, C. (2016) *A Concise Introduction to Decentralized POMDPs*. Springer.
- Papoudakis, G., Christianos, F., Schafer, L. and Albrecht, S.V. (2021) Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. *Proceedings of NeurIPS 2021 Datasets and Benchmarks*.
- Rashid, T., Samvelyan, M., De Witt, C.S., Farquhar, G., Foerster, J. and Whiteson, S. (2018) QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. *Proceedings of ICML 2018*, pp. 4295-4304.
- Samvelyan, M., Rashid, T., De Witt, C.S. et al. (2019) The StarCraft multi-agent challenge. *Proceedings of AAMAS 2019*, pp. 2186-2188.
- Son, K., Kim, D., Kang, W.J., Hostallero, D.E. and Yi, Y. (2019) QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. *Proceedings of ICML 2019*, pp. 5887-5896.
- Sukhbaatar, S., Szlam, A. and Fergus, R. (2016) Learning multiagent communication with backpropagation. *Advances in Neural Information Processing Systems*, 29, pp. 2244-2252.
- Sunehag, P., Lever, G., Gruslys, A. et al. (2018) Value-decomposition networks for cooperative multi-agent learning based on team reward. *Proceedings of AAMAS 2018*, pp. 2085-2087.
- Tan, M. (1993) Multi-agent reinforcement learning: Independent vs. cooperative agents. *Proceedings of ICML 1993*, pp. 330-337.
- Vinyals, O., Babuschkin, I., Czarnecki, W.M. et al. (2019) Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782), pp. 350-354.
- Wei, H., Zheng, G., Gayah, V. and Li, Z. (2021) Recent advances in reinforcement learning for traffic signal control. *ACM SIGKDD Explorations Newsletter*, 22(2), pp. 12-18.
- Yu, C., Velu, A., Vinitisky, E. et al. (2022) The surprising effectiveness of PPO in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35.

Declarations

Funding

This research was supported by the Austrian Science Fund (FWF) under grant P36482 (MultiAgentRL) and the German Federal Ministry of Education and Research (BMBF) under grant 01IS22068 (CooperativeAI). Computational resources provided by the Vienna Scientific Cluster (VSC-4) and HLRN Berlin. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All MARL training code, environment configurations, communication analysis scripts, and evaluation results are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614675. MPE, SMAC, Pommerman, and SUMO are publicly available. Released under MIT License (code) and CC BY 4.0 (results).

Ethical Approval

This study involved computational experiments in simulated multi-agent environments. No human subjects, personal data, or real-world systems were involved. Traffic simulations used synthetic demand patterns. No institutional ethical approval was required.

Appendix A

MARL Method Selection Guide

This appendix provides a decision tree for selecting MARL methods based on interaction structure, agent count, and communication availability.

Step 1: Interaction Structure

- 1a. Fully cooperative (shared reward) --> CTDE methods (QMIX/MAPPO); go to Step 2
- 1b. Fully competitive (zero-sum) --> PBT for robustness; LOLA for exploiting predictable opponents
- 1c. Mixed cooperative-competitive --> MADDPG or QMIX with team decomposition
- 1d. Loosely coupled (mostly independent) --> IQL (simplicity; linear scaling)

Step 2: Communication and Coordination

- 2a. Agents can exchange messages --> CommNet or TarMAC (+8-14% on coordination tasks)
- 2b. No communication channel --> QMIX (implicit coordination through centralised training)
- 2c. Bandwidth-limited communication --> TarMAC with attention (selective messaging)

Step 3: Agent Count and Scalability

- 3a. 2-4 agents --> Any paradigm feasible; choose based on Steps 1-2
- 3b. 5-20 agents --> QMIX (quadratic but manageable); CommNet if communication available
- 3c. 20-100 agents --> Mean-field approximation or IQL with shared parameters
- 3d. 100+ agents --> IQL or hierarchical MARL with team decomposition