Resource Competition in Multi-Agent Reinforcement Learning (MARL)

A Literature Review & Implementation

Anastasia Chernavskaia, Moritz Peist, Nicolas Rauth

June 30, 2025

Outline

Introduction: Why Multi-Agent RL Matters

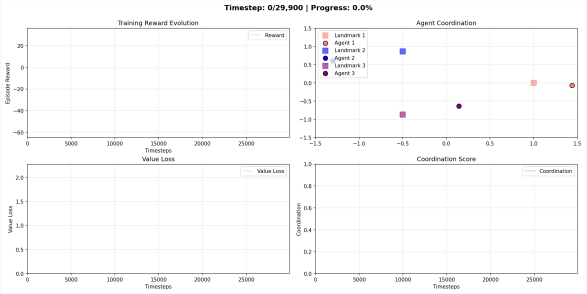
The Multi-Agent Challenge

Traditional RL: One agent learns in a static environment

Multi-Agent RL: Multiple agents learning simultaneously

Key Challenges:

- Non-stationarity (environment changes as agents learn)
- Credit assignment (who contributed to the outcome?)
- Competition vs cooperation
- Scalability with increasing agents



Multi-agent coordination in action

Research Focus & Approach

Project Approach: Literature Review + Code Replication

Focus on resource competition and cooperation dynamics

Literature Survey:

- MADDPG framework
- Value decomposition methods
- Sequential social dilemmas
- Industry applications

Implementation:

- Simple Spread / Adversary environment
- PPO algorithm
- 3 agents coordinating / competing
- Resource allocation challenge

Key Reference: OpenAI's "Learning to cooperate, compete, and communicate" (Lowe, Mordatch, et al., 2017)

Foundational Frameworks

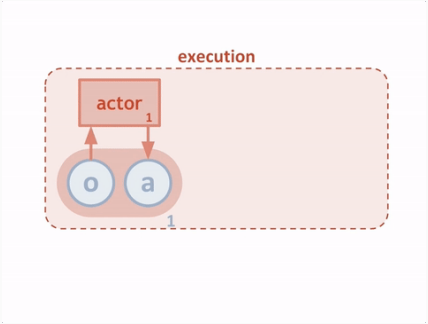
The MADDPG Revolution

Core Innovation: Centralized Training, Decentralized Execution
Train with global information, execute with local observations only

MADDPG Architecture:

- Each agent: own actor $\mu_i(o_i|\theta_i)$
- Centralized critic: $Q_i^{\mu}(x, a_1, \ldots, a_N)$
- Access to all agents' actions during training
- Solves the non-stationarity problem

Key Insight: Environment becomes stationary from critic's perspective (Lowe, Wu, et al., 2017)



MADDPG training architecture

Value Decomposition Methods

Problem: How do we assign credit in team rewards?

QMIX Approach: (Rashid et al., 2020)

- Individual Q-values: $Q_i(o_i, a_i)$
- Team Q-value: $Q_{tot}(s, \mathbf{a})$
- Monotonic mixing: $\frac{\partial Q_{tot}}{\partial Q_i} \geq 0$
- Ensures optimal joint action

QPLEX Extension: (Wang et al., 2021)

- Removes monotonicity constraint
- Uses advantage term
- Better handles negative interactions
- More flexible credit assignment

Key Insight

Value decomposition enables scalable credit assignment while maintaining decentralized execution

COMA: Counterfactual Multi-Agent Policy Gradients

Innovation: Actor-critic with counterfactual baselines (Foerster et al., 2018)

Counterfactual Advantage

$$A_i(s, \mathbf{a}) = Q(s, \mathbf{a}) - \sum_{a_i'} \pi_i(a_i'|o_i) Q(s, (\mathbf{a}_{-i}, a_i'))$$

What this means:

- Compare actual action vs. average over all possible actions
- Isolates agent i's contribution to team performance
- Efficient computation in single forward pass
- Addresses multi-agent credit assignment problem

Applications: Particularly effective in StarCraft unit micromanagement

Emergent Complexity & Competition

Sequential Social Dilemmas

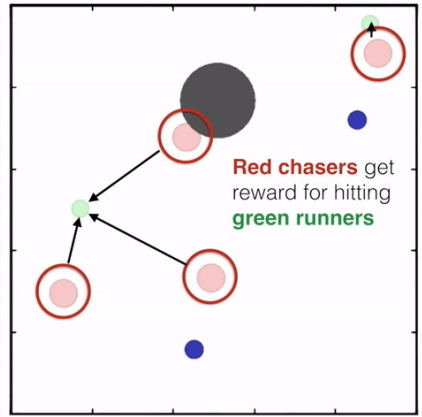
Core Concept: Multi-agent environments with temporal resource competition (Leibo et al., 2017)

Key Properties:

- **Social dilemma:** Individual vs. collective rationality
- **Temporal:** Actions affect future opportunities
- **Resource scarcity:** Limited resources create competition
- **Policy interdependence:** Agents' policies affect each other

Examples:

- Commons Harvest (resource depletion)
- Traffic coordination



Red chasers vs. green runners

Natural Curricula & Autocurricula

Key Insight from OpenAI Research
"Multi-agent environments create natural curricula where difficulty scales with competitor skill"

Autocurriculum Properties:

- **No stable equilibrium:** Continuous pressure for improvement
- **Automatic difficulty scaling:** Environment becomes harder as agents improve
- **Emergent complexity:** Simple rules $\rightarrow$ complex behaviors (Bansal et al., 2018)
- **Robust strategies:** Agents must generalize across opponent types

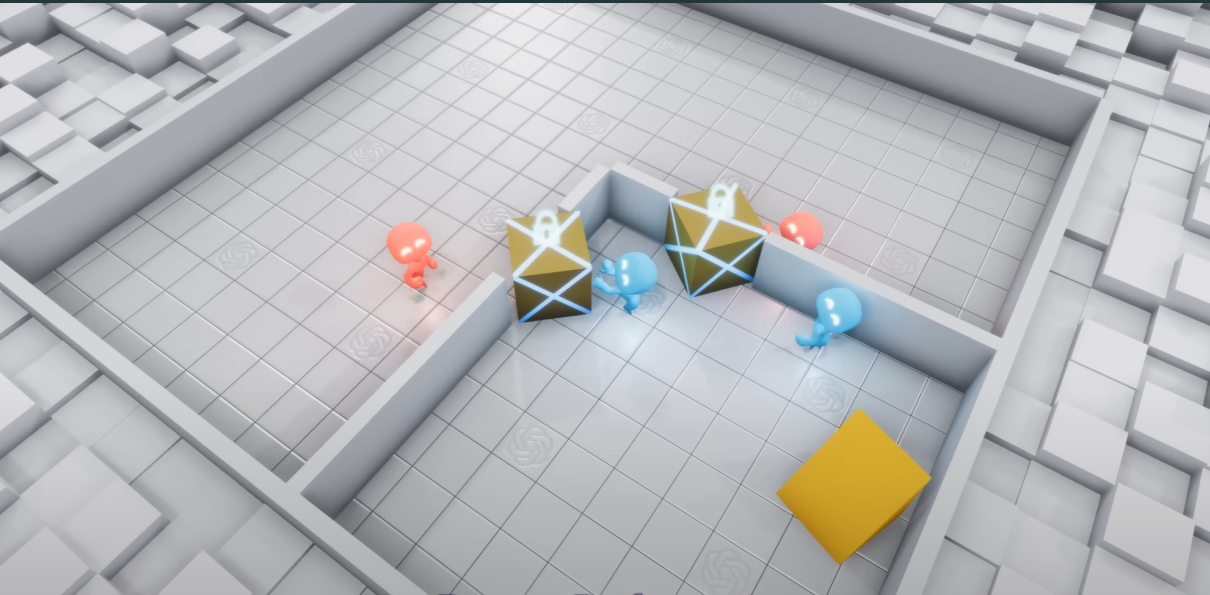
Benefits:

- Self-generated training data
- Avoids over-specialization
- Promotes strategic diversity

Challenges:

- Training instability
- Evaluation difficulties
- Strategy cycling

Natural Curricula & Autocurricula



Our Implementation: PPO in Simple Spread

Experimental Setup

Environment: Simple Spread

- 3 agents, 3 landmarks
- Goal: Cover all landmarks
- Avoid collisions
- Partial observability

Resource Competition:

- Spatial positions as contested resources
- Coordination vs. competition trade-off
- Credit assignment challenge

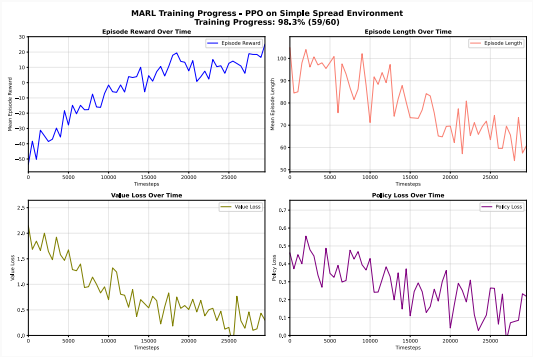
Algorithm: PPO

- Centralized training
- Decentralized execution
- Shared critic network
- Individual policy networks

Why PPO?

- Surprisingly effective in cooperative MARL (Yu et al., 2022)
- Stable training
- Minimal hyperparameter tuning

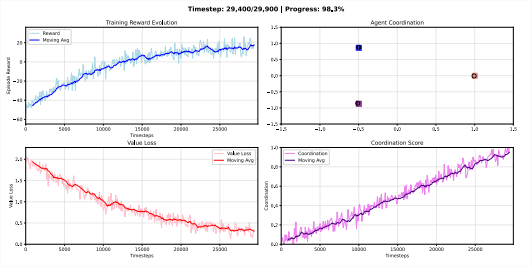
Training Results: Learning Curves



Key Observations:

- **Episode Rewards:** Convergence from -50 to +20 over 30k timesteps
- **Episode Length:** Decreasing length indicates faster coordination
- **Loss Convergence:** Both value and policy losses stabilize

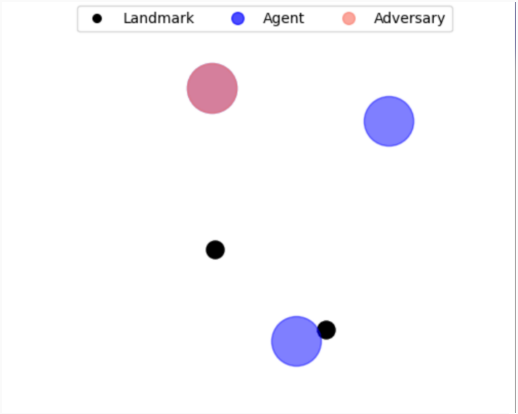
Training Results: Coordination Dashboard



Coordination Development:

- **Reward progression:** Steady improvement in team performance
- **Agent behaviors:** Individual policies learning complementary strategies
- **Coordination score:** Emergent cooperation without explicit communication

Experimental Setup: Simple Adversary



Good Agents Learn:

- Sophisticated movement to avoid capture
- Strategic target selection
- Time movements to reduce exposure

Adversary Agents Learn:

- Predicting good agent movements
- Coordinated chasing behaviors
- Driving agents toward capture zones
- Preventing access to landmarks

Industry Applications

Cloud Computing & Network Management

Resource Allocation Challenges:

- CPU, memory, bandwidth allocation
- Load balancing across servers
- Network routing optimization
- Energy efficiency goals

MARL Applications: (Zhang et al., 2025)

- Autonomous resource managers
- Distributed decision making
- Real-time adaptation
- Scalable to large systems

Benefits:

- Reduced latency
- Better resource utilization
- Fault tolerance
- Adaptive scaling

Challenges:

- High-dimensional state spaces
- Safety constraints
- Real-time requirements
- Partial observability

Multi-Agent Trading Systems: (Shavandi & Khedmati, 2022)

Applications:

- Algorithmic trading strategies
- Market making optimization
- Portfolio management
- Risk assessment systems

Resource Competition:

- Limited market liquidity
- Information asymmetries
- Execution timing
- Capital allocation

MARL Advantages:

- Adaptive to market regimes
- Strategic interaction modeling
- Emergent market behaviors
- Robust to opponent strategies

Considerations:

- Regulatory constraints
- Market impact
- Systemic risk
- Ethical concerns

Current Challenges & Future Directions

Scalability & Robustness Challenges

The Scalability Problem: Most MARL algorithms struggle with increasing # agent

Technical Challenges:

- Exponential action/state spaces
- Communication overhead
- Training instability
- Sample efficiency (Liu et al., 2024)

Current Approaches:

- Hierarchical decomposition
- Graph neural networks
- Attention mechanisms
- Population-based training

Robustness Issues:

- Overfitting to training opponents
- Brittleness to environment changes
- Strategy exploitation
- Evaluation difficulties

Research Directions:

- Domain randomization
- Meta-learning approaches
- Diverse opponent training
- Formal verification methods

Theoretical Foundations & Evaluation

Missing Theoretical Understanding: (Fish et al., 2025)

Theory Gaps:

- Convergence guarantees in multi-agent settings
- Sample complexity bounds
- Equilibrium concepts
- Optimality conditions

Evaluation Challenges:

- No standard benchmarks
- Opponent selection bias
- Metric interpretation
- Reproducibility issues

Emerging Solutions:

- MALIB framework (Zhou et al., 2023)
- Standardized evaluation protocols
- Game-theoretic analysis tools
- Open-source benchmarks

Future Needs:

- Unified theoretical framework
- Better evaluation metrics
- Reproducible research

Conclusion

Literature Review Insights

From Literature:

- **Natural curricula:** Competition creates automatic difficulty scaling
- **No stable equilibrium:** Continuous pressure for improvement
- **Centralized critics:** Enable stable learning in non-stationary environments
- **Value decomposition:** Solves credit assignment while maintaining decentralization
- **Emergent complexity:** Simple individual policies $\rightarrow$ complex team behaviors

From Implementation:

- **PPO effectiveness:** Surprisingly strong in cooperative multi-agent settings
- **Coordination strategies:** Develop through individual learning processes
- **Training dynamics:** Convergence patterns reflect coordination development

Conclusion: The Future of Multi-Agent RL

Bottom Line
MARL represented a paradigm shift toward more realistic, interactive AI systems

What We've Learned:

- Multi-agent environments create unique learning challenges and opportunities
- Algorithmic innovations (MADDPG, QMIX, COMA) address core MARL problems
- Simple algorithms (PPO) can be surprisingly effective with proper implementation
- Real-world applications span critical infrastructure and economic systems

Multi-agent RL: Where competition meets cooperation, and "intelligence" emerges

References

References i

Bansal, T., Pachocki, J., Sidor, S., Sutskever, I., & Mordatch, I. (2018, March 14). **Emergent complexity via multi-agent competition.** https://doi.org/10.48550/arXiv.1710.03748

Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2025).**Algorithmic collusion by large language models.** *AEA Paper Session: AI-Driven Market Dynamics*. https://doi.org/10.48550/arXiv.2404.00806

Foerster, J. N., Farquhar, G., Afouras, T., Nardelli, N., & Whiteson, S. (2018).**Counterfactual multi-agent policy gradients.** *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, 2974–2982.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017).**Multi-agent reinforcement learning in sequential social dilemmas.** *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 464–473.

References ii

Liu, D., Ren, F., Yan, J., Su, G., Gu, W., & Kato, S. (2024).**Scaling up multi-agent reinforcement learning: An extensive survey on scalability issues.** *IEEE Access*, *12*, 94610–94631. https://doi.org/10.1109/ACCESS.2024.3410318

Lowe, R., Mordatch, I., Abbeel, P., Wu, Y., Tamar, A., & Harb, J. (2017, June 7). **Learning to cooperate, compete, and communicate.** Retrieved June 23, 2025, from https://openai.com/index/learning-to-cooperate-compete-and-communicate/

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017).**Multi-agent actor-critic for mixed cooperative-competitive environments.** *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6382–6393.

Rashid, T., Samvelyan, M., De Witt, C. S., Farquhar, G., Foerster, J., & Whiteson, S. (2020).**Monotonic value function factorisation for deep multi-agent reinforcement learning.** *J. Mach. Learn. Res.*, *21*(1), 178:7234–178:7284.

References iii

- Shavandi, A., & Khedmati, M. (2022). **A multi-agent deep reinforcement learning framework for algorithmic trading in financial markets.** *Expert Systems with Applications*, *208*, 118124. https://doi.org/10.1016/j.eswa.2022.118124
- Wang, J., Ren, Z., Liu, T., Yu, Y., & Zhang, C. (2021, October 4). **QPLEX: Duplex dueling multi-agent q-learning.** https://doi.org/10.48550/arXiv.2008.01062
- Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., & Wu, Y. (2022). **The surprising effectiveness of PPO in cooperative multi-agent games.** *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 24611–24624.
- Zhang, J., Liu, Z., Zhu, Y., Shi, E., Xu, B., Yuen, C., Niyato, D., Debbah, M., Jin, S., Ai, B., Xuemin, & Shen. (2025, February 9). **Multi-agent reinforcement learning in wireless distributed networks for 6g.** https://doi.org/10.48550/arXiv.2502.05812

References iv

Zhou, M., Wan, Z., Wang, H., Wen, M., Wu, R., Wen, Y., Yang, Y., Yu, Y., Wang, J., & Zhang, W. (2023).**MALib: A parallel framework for population-based multi-agent reinforcement learning.** *J. Mach. Learn. Res.*, *24*(1), 150:7205–150:7216.