# Multi-Agent Reinforcement Learning and its practical applications in games: a review of the literature

Sean Marco Bonifacio  
University of Northampton

## Abstract

Multi-agent reinforcement learning (MARL) has achieved tremendous success in various artificial intelligence domains such as the gaming industry. This literature review explores the theoretical foundations, principles, algorithms, applications, challenges, and recent advancements in MARL.

Algorithms such as Q-Learning, Policy-based algorithms, SARSA, and more algorithms have laid the groundwork for MARL by providing approaches to agent decision-making and learning. Q-decomposition methods have also further enhanced MARL by decomposing the global value function into local components, allowing for efficient coordination, and learning among agents (Russel & Zimdars, 2003).

The architectures employed in MARL range from centralised to decentralised, with each influencing the coordination and decision-making processes of agents in the environment. Centralised architectures, in which all agents share a single state representation and make collaborative decisions, provide benefits in terms of global coordination and information exchange. Decentralised architectures, on the other hand, encourage agent autonomy by allowing each agent to have its own state representations and decision-making procedures. These designs, when combined with communication and coordination mechanisms, enables MARL systems to model complex interactions and achieve emergent behaviours.

Beyond the gaming industry, MARL is extensively used in a variety of domains, including multi-robot systems, autonomous vehicles, and social dilemmas. Obstacles such as credit assignment, scalability, and ethical considerations persist, emphasising the importance of continued research and innovation in the field.

It is hoped that this paper will provide valuable insights for game developers and researchers regarding MARL and its challenges and advancements.

## Review of Literature

### Overview of MARL

Multi-agent reinforcement learning (MARL) is a subfield of reinforcement learning that focuses on the behaviour of multiple learning agents interacting in a shared environment. MARL algorithms are designed to address various scenarios, including fully cooperative, fully competitive, and mixed cooperative-competitive environments. These algorithms enable agents to learn optimal strategies through interactions with the environment and other agents.

MARL presents unique challenges and applications across different domains. Challenges include nonstationary, varying learning speeds, scalability issues in deep reinforcement learning, and handling partial observability through centralised learning of decentralised policies. To address these challenges, researchers have developed diverse algorithms and methodologies to enhance the performance of MARL systems (Canese, et al., Multi-Agent Reinforcement Learning: A Review of Challenges and Applications, 2021).

The framework of MARL involves models such as Markov Decision Processes (MDPs), Markov Games, Partially-Observable Markov Decision Processes (POMDPs), and Decentralised Partially-Observable Markov Decision Processes (Dec-POMDPs). Single-agent reinforcement learning algorithms like Q-Learning and REINFORCE serve as foundational components in developing MARL algorithms. These algorithms consist of critics for estimating value functions and actors for updating policy parameters based on feedback from the environment. (Canese, et al., Multi-Agent Reinforcement Learning: A Review of Challenges and Applications., 2021).

MARL has been applied in various domains such as robotics, distributed control, telecommunications, and economics due to its ability to address complex tasks that are challenging to solve with preprogrammed behaviours. The field of MARL continues to evolve with advancements in algorithm design, benchmark environments for multi-agent systems, and practical applications across industries (L. Bus¸oniu & Schutter, 2010).

### Frameworks of MARL expanded

#### Markov Decision Processes (MDPs)

Markov Decision Processes (MDPs) serve as a foundational concept in the realm of Multi-Agent Reinforcement Leaning (MARL), providing a mathematical framework for modelling decision-making processes where outcomes are influenced by both random factors and agent actions. In the context of MARL, MDPs extend to game models, enabling the representation of interactions among multiple agents within a shared environment (Albrecht, Christianos, & Schäfer, 2024).

An MDP is defined by a tuple (S, A ,P ,R ,γ), where:

* S represents the set of states,
* A represents the set of actions available to agents,
* P represents the state transitions probabilities,
* R represents the reward function,
* Γ represents the discount factor.

In MARL, MDPs serve as a foundational concept that guides agents in navigating complex decision spaces, learning optimal strategies through interactions with the environment and other agents, and adapting policies to maximise cumulative rewards over time. By leveraging MDPs, researchers can design algorithms that address challenges such as nonstationary, varying learning speeds scalability issues, and partial observability inherent in multi-agent systems.

The formal description of an MDP comprises key components such as states, actions, rewards, and regulations that influence agent decision-making processes. Reinforcement learning algorithms in MARL seek optimal solutions to MDPs by maximising long-term rewards via iterative policy changes and value function estimations.

Researchers can use MDPs in MARL to open up new paths for algorithmic improvement, real-world applications, and a better understanding of multi-agent interactions in complicated contexts.

#### Partially-Observable Markov Decision Processes (POMDPs)

Partially observable Markov Decision Processes (POMDPs) are important in Multi-Agent Reinforcement Learning (MARL) because they allow for principled decision-making in the face of uncertain sensing. In MARL, where agents frequently have restricted sensing capabilities, POMDPs expand the classic Markov Decision Process (MDP) framework to account for circumstances in which agents cannot directly view the underlying state and must rely on observations to make decisions.

In MARL, POMDPs model an agent's relationship with its environment by include elements such as states, actions, transition probabilities, rewards, observations, observation probabilities, and a discount factor. Agents in POMDPs must maintain a belief state based on observations in order to make optimal decisions over time. POMDP applications in MARL include robot navigation difficulties, machine maintenance, and uncertainty-based planning across many domains. (Spaan, 2012)

Researchers created model-based strategies for policy computation and model-free methods for POMDPs within MARL frameworks. POMDPs improve agents' decision-making capabilities in multi-agent systems by tackling difficulties such as incomplete information and unpredictable sensing capabilities. An precise solution to a POMDP offers the optimal action for each feasible belief over global states, maximising expected rewards or minimising costs over an infinite horizon.

The use of POMDPs in MARL improves agent interactions, optimises policies under uncertainty, and overall performance in complex multi-agent systems. Using POMDP principles, researchers may create advanced algorithms that handle real-world difficulties and allow agents to make informed decisions in dynamic and uncertain environments. (Bernstein, Hansen, Amato, & Zilberstein, 2009)

#### Decentralised Partially-Observable Markov Decision Processes

Partially-observable Markov Decision Processes (POMDPs) are extended in the context of Multi-Agent Reinforcement Learning (MARL) to Decentralised Partially-Observable Markov Decision Processes (Dec-POMDPs), which are useful in modelling decision-making processes in multi-agent systems where agents have limited environmental observability. Dec-POMDPs enable a group of agents to collaborate and maximise a global reward based solely on local information, without witnessing a Markovian signal during execution. (Oliehoek)

Dec-POMDPs in MARL simulate interactions between decentralised agents by specifying a tuple consisting of agents, states, actions, transition probabilities, rewards, observations, observation probabilities, and a discount factor. Agents work together to maximise rewards based on local information, making decisions without seeing a full state signal during execution. Dec-POMDPs require developing joint policies that maximise predicted cumulative rewards over a finite horizon.

The use of Dec-POMDPs in MARL improves decision-making capabilities under partial observability, allowing agents to efficiently coordinate actions based on local knowledge. Dec-POMDPs help to optimise agent behaviours and improve overall performance in multi-agent systems by tackling difficulties such as decentralised decision-making and restricted observability. Researchers use the Dec-POMDP framework to create algorithms that address real-world challenges requiring collaborative decision-making under uncertainty. (Learning to Act in Decentralized Partially Observable MDPs, 2018)

#### Single-agent reinforcement learning algorithms

Single-agent reinforcement learning (SARL) methods are used in Multi-Agent Reinforcement Learning (MARL) to improve decision-making and optimise behaviour in multi-agent systems. These techniques, created for individual agents, can be expanded to multi-agent scenarios to promote collaboration, coordination, and efficient learning among several agents interacting in shared environments. (Peng, et al., 2024)

Examples of Single-Agent Reinforcement Learning Algorithms in MARL:

1. Q-Learning: A model-free SARL algorithm that estimates the value of state-action pairs and learns optimal policies through exploration and exploitation.
2. Deep Q-Networks (DQN): An extension of Q-Learning that leverages deep neural networks to approximate Q-values, enabling more complex decision-making in SARL.
3. Policy Gradient Methods: SARL algorithms that directly learn policies without explicit estimating value functions, offering advantages in high-dimensional action spaces and continuous control tasks.
4. Actor-Critic Methods: Hybrid SARL algorithms that combine policy-based and value-based approaches to improve stability and convergence in learning optimal policies.

Reference for examples : (Zhang, Yang, & Başar, 2019).

#### Q-Learning

Q-Learning is a fundamental reinforcement learning method that has been modified and applied to Multi-Agent Reinforcement Learning (MARL) to improve decision-making processes and optimise behaviours in multi-agent systems. Q-Learning is a model-free algorithm that learns optimal policies by balancing exploration and exploitation. In MARL, Q-Learning is a core algorithm that allows agents to learn optimal tactics through interactions with the environment and other agents. Agents can make more informed decisions to maximise cumulative rewards over time by assessing the value of state-action pairings with a Q-table or neural network. Q-Learning algorithms are adaptable to a variety of multi-agent scenarios, including cooperative, competitive, and mixed settings. (Chen, et al., 2018)

In MARL, Q-Learning involves agents updating their Q-values in response to environmental rewards and behaviours. Iterative updates lead agents to optimal strategies that maximise long-term benefits. Deep Q-Networks (DQN), Policy Gradient Methods, and Actor-Critic Methods are extensions of Q-Learning in MARL that provide improved capabilities for managing complex decision spaces and continuous control tasks within multi-agent systems. (Lee, He, Kamalaruban, & Cevher, 2019)

#### Deep Q-Networks (DQN)

Deep Q-Networks (DQN) are a significant improvement in reinforcement learning, especially in the setting of Multi-Agent Reinforcement Learning (MARL). DQN is a technique that uses deep neural networks to approximate the Q-values of state-action pairings, allowing for more complicated decision-making processes in single and multi-agent settings. DQN-based multi-agent systems in MARL are made up for agents who learn and communicate in shared settings. Each agent uses a DQN to calculate Q-values and make decisions based on interactions with its surroundings and other agents. DQN improves agents’ ability to learn complex tactics and optimise behaviours in dynamic and uncertain situations by utilising deep neural networks. (Egorov, 2016)

DQN-based MARL systems provide several advantages over conventional techniques, including simplicity, faster convergence, and improved performance. These systems leverage shared states and incentives while retaining agent-specific actions for updating experience replay pools. Using DQN and Double Q-learning on tasks such as Cartpole-v1, LunarLander-v2, and Maze Traversal, researchers achieved improved performance compared to baseline approaches, demonstrating the usefulness of DQN in multi-agent environments. (Keecheon, 2022)

#### Policy Gradient Methods

Policy gradient methods are a type of reinforcement learning algorithm that directly learns policies without explicitly estimating value functions. Policy Gradient approaches are important in Multi-Agent Reinforcement Learning (MARL) because they allow agents to optimise behaviours, adapt strategies, and collaborate effectively in shared settings. (Fu, Yu, Xu, Yang, & Wu, 2022)

Policy Gradient methods in MARL focus on learning policies that directly map observations to actions, allowing agents to enhance decision-making processes depending on rewards received from the environment. These methods are especially useful in cases where the action space is high-dimensional or continuous since they improve the management of complex decision spaces and the efficient optimisation of policies. In MARL applications, Policy Gradient methods improve agent interactions by allowing them to learn from their experiences, adapt to changing surroundings, and effectively collaborate with others. By adding aspects like intrinsic rewards, self-imitation learning, and enhanced exploration strategies, researchers can create advanced algorithms that increase cooperative behaviours, optimise policy learning, and improve learning efficiency inside multi-agent systems. (Zhao, Chang, & Zhang, 2024)

#### Actor-Critic Methods

Actor-Critic methods are a prominent type of reinforcement learning algorithm that combines the advantages of policy-based and value-based approaches. In the context of Multi-Agent Reinforcement Learning (MARL), extending Actor-Critic methods to multi-agent systems entails tackling communication, coordination, and efficient learning issues in complicated contexts.

Key aspects of Policy Gradient Methods in MARL:

1. On-Policy Learning
2. Suitability for Collaborative Settings
3. Enhanced Exploration Techniques
4. Efficient Learning in Complex Techniques

### Key Challenges in Multi-Agent Reinforcement Learning

Multi-agent reinforcement learning (MARL) presents several key challenges that must be addressed for the successful development and deployment of efficient algorithms. One of the primary challenges is nonstationary, which refers to the dynamic nature of the environment due to interactions among multiple agents. As highlighted by (Canese, et al, 2021), in MARL, each agent perceives other agents as part of the environment, leading to constant changes influenced by all agents’ actions. This introduces a significant hurdle in achieving convergence and maintaining acceptable performance, especially when a large number of agents are involved. Additionally, scalability is another crucial challenge as developing algorithms capable of accomodating real-world problems involving numerous agents requires careful consideration. Futhermore, observaility poses a challenge in terms of acquiring complete information about the state and actions of other agents within the environment, thereby impacting decision-making processes. The scalability to a high number of agents is an essential feature that must be taken into account when developing algorithms that can be applied to real-world problems. MARL has shown potential applications in various real-world scenarios. One such application is in robotics, where MARL enables multiple agents to collaborate and learn from each other to accomplish complex tasks efficiently and effectively (Canese, et al., Multi-Agent Reinforcement Learning: A Review of Challenges and Applications, 2021). The versatality of MARL makes it well-suited for addressing diverse challenges across different domains by harnessing the collective intelligence and coordination capabilities of multiple autonomous entities.

# References

Albrecht, S. V., Christianos, F., & Schäfer, L. (2024). *Multi-Agent Reinforcement Learning.* Cambridge, Massachusetts: MIT Press.

Bernstein, D. S., Hansen, E. A., Amato, C., & Zilberstein, S. (2009). Policy Iteration for Decentralized Control of Markov Decision Processes. *Journal of Artificial Intelligence Research 34*, 89-132.

Canese, L., Cardarilli, G., Di Nunzio, L., Fazzolari, R., Giardino, D., Re, M., & Spanò, S. (2021). Multi-Agent Reinforcement Learning: A Review of Challenges and Applications. *Applied Sciences* .

Canese, L., Cardarilli, G., Di Nunzio, L., Fazzolari, R., Giardino, D., Re, M., & Spanò, S. (2021). Multi-Agent Reinforcement Learning: A Review of Challenges and Applications. *Applied Science* .

Canese, L., Cardarilli, G., Di Nunzio, L., Fazzolari, R., Giardino, D., Re, M., & Spanò, S. (2021). Multi-Agent Reinforcement Learning: A Review of Challenges and Applications. *Applied Sciences*.

Chen, Y., Zhou, M., Wen, Y., Yang, Y., Su, Y., Zhang, W., . . . Liu, H. (2018). Factorized Q-learning for large-scale multi-agent systems. *Proceedings of the First International Conference on Distributed Artificial Intelligence .* International Conference on Distributed Artificial Intelligence .

Egorov, M. (2016). *Multi-Agent Deep Reinforcement Learning.* Stanford, CA.

Fu, W., Yu, C., Xu, Z., Yang, J., & Wu, Y. (2022, July 10). *Why do Policy Gradient Methods work so well in Cooperative MARL? Evidence from Policy Representation*. Retrieved from Bair Barkeley Artificial Intelligence Research: https://bair.berkeley.edu/blog/2022/07/10/pg-ar/

Keecheon, K. (2022). Multi-Agent Deep Q Network to Enhance the Reinforcement Learning for Delayed Reward System. *Applied Sciences*.

L. Bus¸oniu, R. B., & Schutter, B. D. (2010). Multi-agent reinforcement learning: An Overview. In D. Srinivasan, & L. Jain, *Innovations in Multi-Agent Systems and Applications* (pp. 183-221). Berlin, Germany: Springer.

Learning to Act in Decentralized Partially Observable MDPs. (2018). In J. Dibangoye, & O. Buffet, *Proceedings of the 35th International Conference on Machine Learning* (pp. 1233-1242). PMLR.

Lee, D.-h., He, N., Kamalaruban, P., & Cevher, V. (2019). Optimization for Reinforcement Learning: From a single agent to cooperative agents. *IEEE Signal Processing Magazine 37*, 123-135.

Oliehoek, F. A. (n.d.). Decentralized POMDPs.

Peng, S., Xiong, G., Yang, J., Shen, Z., Tamir, T., Tao, Z., . . . Wang, F.-Y. (2024). Multi-Agent Reinforcement Learning for Extended Flexible Job Shop Scheduling. *Machines*.

Russel, S., & Zimdars, A. L. (2003). Q-Decomposition for Reinforcement Learning Agents. *Proceedings of the Twentieth Conference on Machine Learning* .

Spaan, M. T. (2012). Partially Observable Markov Decision Processes. In M. T. Spaan, *Reinforcement Learning. Adaptation, Learning, and Optimization* (pp. 387-414). Berlin: Springer.

Zhang, K., Yang, Z., & Başar, T. (2019). *Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms.* arXiv.

Zhao, L., Chang, T., & Zhang, L. (2024). Multi-agent cooperation policy gradient method based on enhanced exploration for cooperative tasks. *International Journal of Machine Learning and Cybernetics*, 1431-1452.