Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Review:

Optimization methods in fully cooperative scenarios: a review of multiagent reinforcement learning*

Tao YANG[†§1,2], Xinhao SHI[†§1,2], Qinghan ZENG[2], Yulin YANG[2], Cheng XU[‡1], Hongzhe LIU[1]

[1]*Beijing Key Laboratory of Information Service,Beijing Union University, Beijing 100101, China*

[2]*Science and Technology Innovation Research Center of ARI, Unit 32178 of the PLA, Beijing 100012, China*

[†]E-mail: 20231083510923@buu.edu.cn; 20221083510927@buu.edu.cn; xc-f4@163.com

Received Apr. 6, 2024; Revision accepted Sept. 6, 2024; Crosschecked

Abstract: Multiagent reinforcement learning has become a dazzling new star in the field of reinforcement learning in recent years, demonstrating its immense potential across a plethora of application scenarios. The reward function directs agents to explore and make optimal decisions within their environments by establishing evaluation criteria and feedback mechanisms. Concurrently, cooperative objectives at the macro level provide a trajectory for agents' learning, ensuring alignment between individual behavioral strategies and the overarching system goals. The interplay between reward structures and cooperative objectives not only bolsters the effectiveness of individual agents but also fosters interagent collaboration, offering both momentum and direction for the development of swarm intelligence and the harmonious operation of multiagent systems. This review delves deeply into the methods for designing reward structures and optimizing cooperative objectives in multiagent reinforcement learning, along with the most recent scientific advancements in these fields. The article meticulously reviews the application of simulation environments in cooperative scenarios and discusses future trends and potential research directions in the field, providing a forward-looking perspective and inspiration for subsequent research efforts.

Unedited

Key words: Multiagent reinforcement learning; Cooperative framework; Reward function; Cooperative objective optimization

https://doi.org/10.1631/FITEE.2400259 **CLC number:**

1 Introduction

Multiagent reinforcement learning (MARL) has become a research hotspot in the field of RL in recent years and has made significant progress, opening up a series of highly complex and challenging application areas, such as autonomous driving (Zhang KQ et al., 2021; Huang et al., 2024; Ren FY et al., 2024; Ren Y et al., 2024), drone collaboration (Jia et al., 2023; Nian et al., 2024; Wang BL et al., 2024), smart cities (Wu T et al., 2020; Qiao et al., 2024), smart grids (Xu et al., 2020; Wang JH et al., 2021a; Gou et al., 2022), and robotic cooperation (Chen HB et al., 2023; Gu et al., 2023).

The reward function is a fundamental component of MARL, which guides agents to learn how to make optimal decisions within an environment by defining and providing a feedback mechanism (Icarte et al., 2022). The reward function operates by assessing the outcomes of agents' actions and determining the appropriate rewards or penalties to be assigned. This feedback mechanism directly influences

‡ Corresponding author
§ These authors contributed equally to this work

* Project supported by the Technology Preliminary Research Project (No. 50911030701), the National Natural Science Foundation of China (Nos. 62171042 and 62102033), the R&D Program of Beijing Municipal Education Commission (No. KZ202211417048), the Project of Construction and Support for High-Level Innovative Teams of Beijing Municipal Institutions (No. BPHR20220121), the Beijing Natural Science Foundation (Nos. 4232026 and 4242020), the Academic Research Projects of Beijing Union University (No. ZKZD202302)

© Zhejiang University Press 2024

the agents' behavioral strategies, as they learn to prioritize actions that yield higher rewards and avoid those that lead to penalties. In the domain of multiagent systems (MASs), crafting an appropriate reward structure is paramount, as it necessitates judicious reward allocation among individual agents and collective entities within the team. This is done to foster various interactive dynamics, such as cooperation or competition, and to guarantee the alignment of agent objectives toward the attainment of shared goals (Shou and Di, 2020). A meticulously formulated reward function is instrumental in augmenting the efficiency of collective learning processes, optimizing synergistic collaboration among agents, and equipping the agents with the capacity for adaptation and proficiency in environments characterized by their complexity and dynamism.

The optimization of cooperative objectives plays a crucial role within the framework of MARL, as it integrates and optimizes individual reward signals, providing a quantified optimization criterion for the adjustment of agent strategies. By aiming for the maximization of long-term cumulative rewards, it guides agents to search for the optimal sequence of actions within the strategy space and ensures the alignment of the agents' decision-making processes with collective goals (Yang NK et al., 2023). This optimization not only reinforces the goal-directedness of agents but also promotes efficient collaboration and the development of adaptability within MASs.

The reward structure maintains a complementary and synergistic relationship with the objective function, in which the reward structure forms the cornerstone of cooperative goal optimization, providing immediate and micro-level guidance for the behavior of agents. Complementarily, the cooperative goals establish the direction of the agents' learning process at the macro level, ensuring the consistency of the agents' behavioral learning strategies with the overall goals of the system (Rădulescu et al., 2020). This bidirectional interaction not only optimizes the performance of individual agents but also promotes collaboration among agents, providing momentum and direction for the emergence of collective intelligence and the harmonious operation of MASs. Therefore, considering the importance of reward function design and cooperative goal optimization in MARL, this paper provides a comprehensive review of the current research status in related fields. We systematically summarize the representative works in the domains of reward shaping and cooperative goal optimization, analyzing the characteristics, advantages, and limitations of different methods.

Some review works have been conducted on MARL, encompassing aspects such as MAS (Wang JR et al., 2022), cooperative MARL (Oroojlooy and Hajinezhad, 2023; Zhao LY et al., 2023), MARL based on transfer learning, communication-based MARL (Khan R et al., 2023), causality-based MARL, as well as game theory-based RL. Additionally, some reviews address specific challenges, such as dealing with nonstationarity (Papoudakis et al., 2019), balancing the exploration-exploitation trade-off (Hao JY et al., 2024), or addressing issues of scalability in MARL (Yuan WL et al., 2024).

2 Multi-agent reinforcemen learning

Although there has been extensive research in the field of MARL, a systematic review of the methods and progress in reward structure design and cooperative goal optimization in fully cooperative scenarios has not yet been conducted. Considering the core role of reward structures and cooperative goal optimization in promoting effective cooperation among agents, this paper aims to comprehensively review the key concepts and latest research achievements in this domain. We hope to provide researchers with a panoramic reference through this article, to foster a deeper understanding of rewards and optimization strategies in fully cooperative MASs, and to encourage further development of related technologies.

The structure of this review is presented in Fig. 1. The second section delves into the fundamental theories of multi-agent reinforcement learning, including Markov decision processes, partially observable Markov decision processes, the centralized training with decentralized execution paradigm, and game theory under imperfect information. The third section elaborates on the design principles and practices of reward structures, focusing on the stimulation of intrinsic motivations, reshaping of reward functions, and the distribution mechanisms for team-level rewards, and introduces related research methods. The fourth section turns to how to optimize cooperative goals, discussing in detail the application of

Unedited

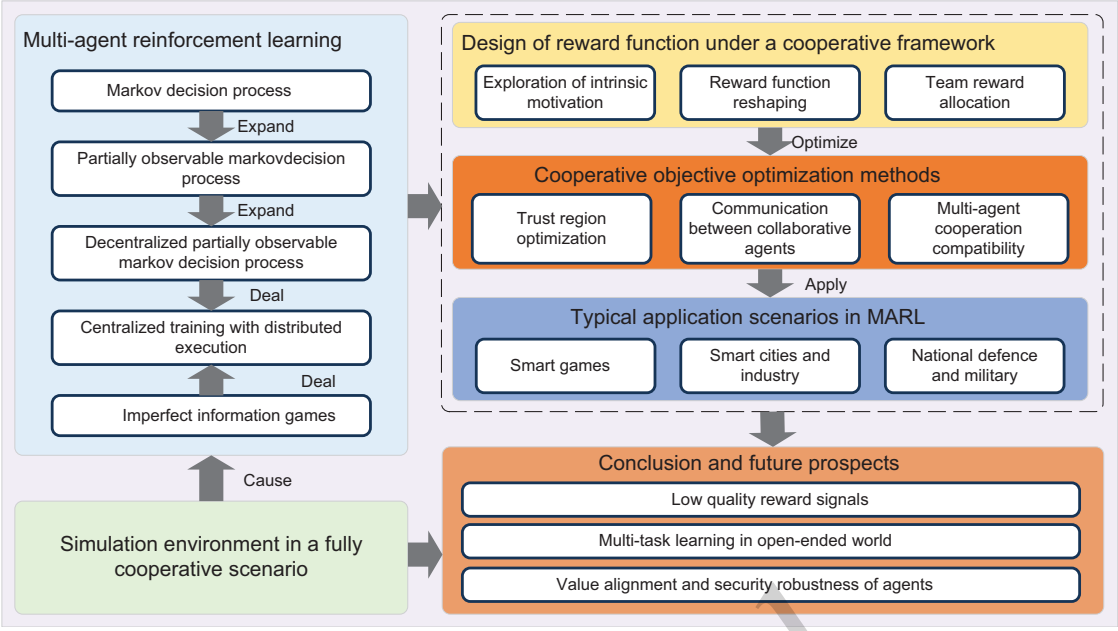

Fig. 1 Overall structure of the review

Unedited

trust region methods, communication mechanisms, and analysis and optimization strategies of cooperative compatibility in the optimization of cooperative goals. The fifth section sorts out and comments on the existing mainstream simulation environments. The article concludes not only by summarizing the content but also by providing a forward-looking outlook on future research directions, hoping to attract more research interest and guide future work.

2.1 Markov decision process (MDP)

The classic definition of an MDP (White III and White, 1989) is given by a five-tuple (S, A, P, R, γ) structure, which can be formally defined as follows: S is a finite set of states, representing all possible configurations that the environment can be in. A is a finite set of actions available to the agent. P is the state transition probability function, $P(s_{t+1}|s_t, a_t)$, which gives the probability of moving to state s_{t+1} from state s_t after taking action a_t. R is the reward function, $R(s_t, a_t, s_{t+1})$, which defines the immediate reward received after taking action a_t in state s_t and landing in state s_{t+1}. γ is the discount factor, $\gamma \in [0, 1]$, used to weigh the importance of future rewards compared to immediate rewards.

In this framework, an agent interacts with the environment in a sequence of discrete timesteps. At each timestep t, the agent perceives the current state s_t of the environment. Based on the current state and its policy π, which is a function mapping states to actions, the agent selects an action a_t. After executing this action, the environment transitions to a new state s_{t+1} according to the transition probabilities $P(s_{t+1}|s_t, a_t)$. The agent receives a reward r_{t+1} as defined by the reward function $R(s_t, a_t, s_{t+1})$. If the agent is learning, it updates its policy based on the received reward and the newly observed state.

In RL (Andrew, 1999), the objective of an agent in an MDP is gcnerally to maximize the expected sum of discounted rewards over time. This objective is often formalized as the return, which is the cumulative reward received over the course of an episode or interaction with the environment:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R(s_{t+k}, a_{t+k}, s_{t+k+1}), \tag{1}$$

where G_t is the total discounted reward from time step t, and γ is the discount factor that balances the importance of immediate versus future rewards. The goal is to find a policy π^* that maximizes the expected return from each state over all possible policies. To assess the quality of a policy π, value functions are used. The "state-value function", $V^\pi(s)$, represents the expected return when starting in state s and following policy π thereafter:

$$V^\pi(s) = \mathbb{E}[G_t|S_t = s]. \tag{2}$$

The "action-value function", also known as the Q-function, $Q^\pi(s,a)$, represents the expected return after taking an action a in state s under policy π:

$$Q^\pi(s,a)=\mathbb{E}[G_t|S_t=s,A_t=a]. \tag{3}$$

In Q-learning, an off-policy algorithm, the agent learns an approximation of the action-value function, which is independent of the policy being followed. The Q-learning algorithm updates the Q-values using the Bellman equation as follows:

$$\begin{aligned}Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha[R_{t+1}\\ +\gamma\max_a Q(s_{t+1},a)-Q(s_t,a_t)],\end{aligned} \tag{4}$$

where α is the learning rate. Temporal Difference (TD) learning algorithms (Sutton, 1988) are a family of model-free reinforcement learning methods that learn directly from raw experience without a model of the environment's dynamics. TD methods are based on the idea that the estimate of the value function at one time step can be updated towards the estimate of the value function at the next time step. The simplest TD method, TD(0), updates the value function estimate $V(s)$ for the non-terminal states as follows:

$$V(s_t)\leftarrow V(s_t)+\alpha[R(s_{t+1})+\gamma V(s_{t+1})-V(s_t)], \tag{5}$$

where $R(s_{t+1})$ is the reward received after transitioning to the next state s_{t+1}, and γ is the discount factor. The TD error, $R(s_{t+1})+\gamma V(s_{t+1})-V(s_t)$, represents the difference between the estimated value of the current state and the better estimate provided by the subsequent state and reward. TD learning can also be used to update Q-values in a similar fashion, leading to algorithms like SARSA (State-Action-Reward-State-Action) (Singh et al., 2000), where the update is done using the action actually taken, as opposed to the max over possible actions used in Q-learning. Theoretically, given the Bellman optimal operator $\mathcal{B}^*$, the solution process of the Q-value function can be defined as:

$$\begin{aligned}&(\mathcal{B}^*Q)(s,a)\\ &=\sum_{s'}P(s'|s,a)\left[R(s,a)+\gamma\max_{a'\in\mathcal{A}}Q(s',a')\right],\end{aligned} \tag{6}$$

where $P(s'|s,a)$ is the transition probability from state s to state s' under action a, $R(s,a)$ is the reward function, and γ is the discount factor. In order to establish the optimal Q-value function, the Q-learning algorithm uses the fixed point iteration of the Bellman equation to solve the Q-value function:

$$Q^*(s,a)=(\mathcal{B}^*Q^*)(s,a), \tag{7}$$

which has a unique solution. In practice, when the environment model is unknown, the state and action spaces are discrete, and all actions can be repeatedly sampled in all states, the above Q-learning method can be guaranteed to converge to the optimal solution.

Policy gradient methods are a class of reinforcement learning algorithms that directly adjust the parameters of the policy based on the gradient of the expected reward with respect to the policy parameters. Unlike value-based methods, which first learn a value function and derive a policy from it, policy gradient methods directly parameterize the policy and update the policy parameters θ by ascending on the expected reward. The objective function $J(\theta)$ for policy parameters θ is typically the expected cumulative reward, and the gradient of this objective with respect to θ is:

$$\nabla_\theta J(\theta)=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t,s_{t+1})\nabla_\theta\log\pi_\theta(a_t|s_t)\right]. \tag{8}$$

However, the gradient will also be subject to the unknown effects of policy changes on the state distribution. Therefore, the researchers derived a solution that does not involve state distribution based on the Policy Gradient theorem:

$$\begin{aligned}&\nabla_\theta V^{\pi_\theta}(s)\\ &=\mathbb{E}_{s\sim\mu^{\pi_\theta}(\cdot),a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta\log\pi_\theta(a|s)\cdot Q^{\pi_\theta}(s,a)\right],\end{aligned} \tag{9}$$

where μ^{π_θ} is the occupancy measure of the policy π_θ, and $\nabla_\theta\log\pi_\theta(a|s)$ is the update score evaluation of the policy. When the policy is deterministic and the action space is continuous, we can further obtain the Deterministic Policy Gradient (DPG) theorem:

$$\begin{aligned}&\nabla_\theta V^{\pi_\theta}(s)\\ &=\mathbb{E}_{s\sim\mu^{\pi_\theta}(\cdot)}\left[\nabla_\theta\pi_\theta(a|s)\cdot\nabla_a Q^{\pi_\theta}(s,a)|a=\pi_\theta(s)\right].\end{aligned} \tag{10}$$

2.2 Partially observable MDP (POMDP)

A POMDP (Bernstein et al., 2002) is an extension of the classic MDP (White III and White,

Unedited

1989) framework to multiagent environments, where multiple agents operate independently and simultaneously, with partial observability of the global state. Decentralized MDPs (DecMDPs) constitute a formal framework for modeling decision-making in cooperative MASs.A Dec-MDP can be defined by the tuple $(\mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, \mathcal{R}, \mathcal{T}, \gamma)$,$\mathcal{N} = \{1, 2, ..., n\}$ is the set of agents indexed by i. S is the set of states of the environment. $\mathcal{A}^i$ is the set of actions available to agent i. The joint action space for all agents is given by the Cartesian product $A = A_1 \times A_2 \times \ldots \times A_N$. $\mathcal{T} : S \times A \times S \to [0, 1]$ is the state transition probability function, which defines the probability of transitioning from one state to another state given the joint action of all agents. $\mathcal{R} : S \times A \to \mathbb{R}$ is the reward function, indicating the immediate reward received by the agents for performing a joint action in a given state. γ is the discount factor, representing the difference in importance between future and present rewards, with $0 \leq \gamma < 1$.

At each time step t, the agents are in a state $s_t \in \mathcal{S}$. Based on their local observations, which may be a partial view of s_t, each agent $i \in \mathcal{N}$ selects an action $a^i \in \mathcal{A}^i$ according to its policy π^i, which maps its observations to actions. The collective actions of all agents $a_t = (a_t^1, a_t^2, ..., a_t^N)$ form a joint action that affects the environment. The state transitions to s_{t+1} with the probability given by the transition function $\mathcal{T}(s_{t+1}|s_t, a_t)$. Concurrently, each agent receives a reward r_t^i based on the reward function $\mathcal{R}^i(s_t, a_t)$ .This interaction pattern continues over the course of the decision-making horizon, creating a trajectory of states and actions that are evaluated for their expected cumulative reward.

In Dec-MDPs, we define several value functions to evaluate the potential of states and actions.

State-Value Function $V_i^\pi(s)$: The value of a state s for an agent i under a joint policy π is the expected return starting from s and following π thereafter.

$$V_i^\pi(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{i,t+k} \mid S_t = s \right]. \tag{11}$$

Action-value function $Q_i^\pi(s, a)$: The value of taking an action a in state s for agent i under a joint policy π is the expected return starting from s, taking action a, and following π thereafter.

$$Q_i^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{i,t+k} \mid S_t = s, A_{i,t} = a \right]. \tag{12}$$

Advantage function $A_i^\pi(s, a)$: The advantage of taking an action a over others in state s for agent i under a joint policy π, defined as the difference between the action value and the state-value.

$$A_i^\pi(s, a) = Q_i^\pi(s, a) - V_i^\pi(s). \tag{13}$$

Each agent aims to maximize its expected cumulative reward over time. This is done often by finding—for each agent—an optimal policy π_i^* that maximizes the expected return.

In a multiagent environment, the actions of each agent not only affect their own return but may also affect that of other agents. Thus, defining a state-action value function that accounts for the actions of all agents is beneficial. For an ordered subset $i_{1:m}$, the multiagent state-action value function $Q_{\boldsymbol{\pi}}^{i_{1:m}}$ is defined as follows:

$$\begin{aligned} &Q_{\boldsymbol{\pi}}^{i_{1:m}}\left(s, \boldsymbol{a}^{i_{1:m}}\right) \\ &\triangleq \mathbb{E}_{\mathbf{a}^{-i_{1:m}} \sim \boldsymbol{\pi}^{-i_{1:m}}} \left[Q_{\boldsymbol{\pi}}\left(s, \boldsymbol{a}^{i_{1:m}}, \mathbf{a}^{-i_{1:m}}\right) \right], \end{aligned} \tag{14}$$

where s denotes the environmental state, $\boldsymbol{a}^{i_{1:m}}$ is the vector of actions for agents in the subset $i_{1:m}$, $\mathbf{a}^{-i_{1:m}}$ is the vector of actions for agents in the complementary set, $\boldsymbol{\pi}$ represents the collection of policies for the agents, and $Q_{\boldsymbol{\pi}}$ is the overall state-action value function under policy $\boldsymbol{\pi}$.

Furthermore, to analyze the interactions among different sets of agents, we can define the multiagent advantage function $A_\pi^{i_{1:m}}$. Assuming that we have two disjoint sets of agents $j_{1:k}$ and $i_{1:m}$, the multiagent advantage function is defined as follows:

$$\begin{aligned} &A_\pi^{i_{1:m}}\left(s, \boldsymbol{a}^{j_{1:k}}, \boldsymbol{a}^{i_{1:m}}\right) \triangleq \\ &Q_\pi^{j_{1:k}, i_{1:m}}\left(s, \boldsymbol{a}^{j_{1:k}}, \boldsymbol{a}^{i_{1:m}}\right)) - Q_\pi^{j_{1:k}}\left(s, \boldsymbol{a}^{j_{1:k}}\right), \end{aligned} \tag{15}$$

where $A_\pi^{i_{1:m}}$ represents the additional value that the actions $\boldsymbol{a}^{i_{1:m}}$ of the subset $i_{1:m}$ contribute, given the actions $\boldsymbol{a}^{j_{1:k}}$ of the subset $j_{1:k}$, in state s over what is expected. This helps us understand how the strategies of different sets of agents influence the overall state-action value function in an MAS.

2.3 Decentralized POMDP (Dec-POMDP)

A Dec-POMDP is a powerful framework for modeling MARL in fully cooperative scenarios.

Unedited

As an extension and generalization of the classic POMDP, Dec-POMDP provides a formal structure to capture the complex interactions and decision-making processes among multiple agents in partially observable environments. The Dec-POMDP is defined by the tuple $< I, S, A, T, R, \Omega, O, \gamma >$, where each component plays a crucial role in describing the MAS.

The set of agents participating in the decision-making process is denoted by I. Each agent $i \in I$ has its own local observation space Ω_i and action space A_i. The global state space S represents the set of all possible states of the system, encompassing the complete information about the environment and the states of all agents. The joint action space $A = \prod_i A_i$ is the Cartesian product of the individual action spaces, where a joint action $a \in A$ is a vector of actions taken by each agent in the set I.

The transition function $T(s, a, s') : S \times A \to \delta(S)$ is a probability distribution that governs the system's dynamics, specifying the probability of transitioning to the next state s' given the current state s and the joint action a. Here, $\delta(S)$ denotes the set of probability distributions over the state space S. The joint observation space $\Omega = \prod_i \Omega_i$ is the Cartesian product of the individual observation spaces, where each agent i receives an observation $o_i \in \Omega_i$ at each time step according to the joint observation probability $O(o, s', a) = P(o|s', a)$. This probability distribution specifies the likelihood of observing o given the next state s' and the joint action a.

The reward function $R(s_t, a_t)$ determines the joint reward r_t received by the agents at each time step t, based on the current state s_t and the joint action a_t. The discount factor $\gamma \in [0, 1]$ is used to balance the importance of immediate and future rewards, enabling the calculation of the expected return.

In Dec-POMDP, each agent i maintains a local observation history $o_t^i = (o_1^i, o_2^i, ..., o_t^i)$, where $o_t^i \in \tilde{\Omega}_t^i$ represents the sequence of observations received by agent i up to time step t. Based on this local observation history, each agent follows a local policy $\pi_i(a_i|o^i)$, which is a probability distribution over actions, specifying the likelihood of agent i choosing action a_i given its observation history o^i.

The Dec-POMDP provides a rich and expressive framework for modeling and solving MARL problems in fully cooperative scenarios. By capturing the partial observability, decentralized decision-making, and joint rewards, Dec-POMDP enables the design and analysis of algorithms that can effectively coordinate the actions of multiple agents to achieve common goals.

2.4 Centralized training with distributed execution

MARL has emerged as a powerful framework for tackling complex environments and tasks that require the coordination and cooperation of multiple agents to achieve a common goal. The choice of training paradigm plays a crucial role in the design and implementation of MARL algorithms, as it determines the level of centralization and decentralization in the training and execution processes. This section delves into the three main training paradigms in MARL: centralized training and centralized execution (CTCE), centralized training and decentralized execution (CTDE), and decentralized training and decentralized execution (DTDE). We explore the characteristics of each paradigm, providing insights into their suitability for different multiagent learning problems.

CTCE is a fully centralized MARL paradigm that addresses the coordination problem among multiple agents by introducing a centralized controller. In both the training and the execution phases, this centralized controller possesses complete information, including the states, observations, and rewards of all agents. It is responsible for learning a centralized joint policy π, which maps the joint states of all agents to the joint action space, and guides the action selection of the agents during the execution phase.

DTDE is a fully distributed MARL paradigm. The main idea is that during both the training and the execution phases, each agent operates completely independently, with no coordination or communication. Each agent i has its own independent policy π_i, which maps local observations o_i to the action space A_i. Throughout the training and execution phases, agents rely entirely on their own local information for learning and decision-making, without any communication or information exchange.

As shown in Fig. 2, CTDE is a training paradigm that allows for information exchange and communication among agents during the training phase, thereby enabling them to gain additional in-

Unedited

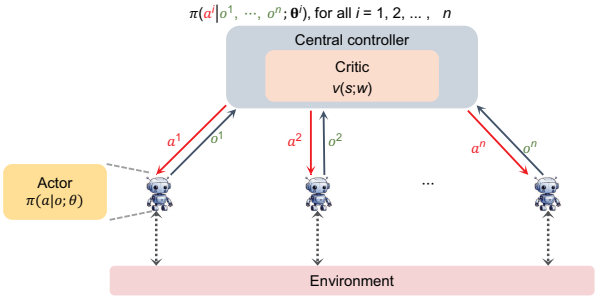

Fig. 2 During training, agents collaboratively learn with access to global state information using a centralized method. During execution, agents independently make decisions using only their local information

formation to learn better strategies. This information exchange can involve the sharing of observations, actions, rewards, and other types of information. By sharing information, agents can better understand the environment and the actions of other agents and adjust their strategies accordingly. The CTDE approach consists of two key components:

2.4.1 Centralized training

During the training phase, a central entity has access to the full state of the environment as well as the actions and observations of all agents. This central entity can be a computational process that aggregates and processes the global information to guide the learning of each agent. Centralized training allows for the use of global state information to perform credit assignment more accurately, as the impact of each agent's actions on the global outcome can be better assessed.

2.4.2 Decentralized execution

Once trained, the agents execute their learned policies independently, without access to the global state or the actions of other agents. Each agent's policy is conditioned solely on its local observation history, making the execution feasible in environments where in communication constraints or privacy concerns preclude the sharing of global information.

2.5 Imperfect information games

In the realm of RL, autonomous agents iteratively refine their behavioral policies by embarking on exploratory actions and assimilating the consequential reward signals emitted by the environment. This iterative learning paradigm draws its theoretical underpinnings from the framework of MDPs, wherein the decision-making process of an agent is predicated exclusively upon the immediate state of the environment, disregarding historical states or future considerations. Nonetheless, the complexity of the real world frequently encompasses scenarios wherein multiple agents coexist and interact, necessitating collective decision-making. An agent's action within such an MAS not only influences its own reward prospects but also affects the reward dynamics and strategic selections of its peers, thereby engendering a situation reminiscent of game-theoretical constructs. In this context, game theory burgeons as an indispensable analytical tool, providing the necessary conceptual apparatus to dissect and understand the strategic interplays within MASs. Particularly, in fully cooperative scenarios, wherein agents strive towards a common goal, the orchestration of reward structures and the optimization of collaborative strategies are of paramount importance.

In cooperative MASs, the most common resolution is to achieve a Nash equilibrium. Nash equilibrium refers to a joint strategy $\pi^{\mathrm{NE}} \in \Pi$, in which no agent can unilaterally increase the joint reward.

Imperfect information games are an important branch of game theory, relaxing the assumption of perfect information in traditional game models and allowing players to face incomplete information during the game. To address this situation, the Bayesian game model (Harsanyi, 1967) takes into account the types of players and their prior distributions, allowing players to infer the distribution of other players' types based on their own type. An n-person Bayesian game can be represented as a quintuple (Gibbons and Gibbons, 1992):

$$G_b = (N, \{A_i\}_{i\in N}, \{\Theta_i\}_{i\in N}, \{\rho_i\}_{i\in N}, \{\gamma_i\}_{i\in N}), \tag{16}$$

where N is the set of players, A_i is the set of actions for player i, Θ_i is the set of types for player i, ρ_i is the prior distribution of player i over the types of the other players, and γ_i is the payoff function for player i, which depends on the type.

3 Design of reward function under a cooperative framework

In addressing complex real-world challenges, especially those necessitating coordination and coop-

Unedited

eration among autonomous entities, fully MASs have emerged as a pivotal area of research.

In this section, we scrutinize the extant research on reward structure design within a cooperative multiagent context. We commence with an examination of the foundational tenets of reward design and their ramifications for agent behavior. Furthermore, we scrutinize the intricacies of reward shaping in fostering effective cooperation, addressing issues such as the credit assignment problem, the alignment of incentives, and the nuances of reward modification.

3.1 Intrinsic motivation exploration (IMOE)

In the human brain, dopamine is a neurotransmitter that plays a significant role in the reward process. The release of dopamine is associated with stimuli and behaviors related to rewards. When we experience a sense of reward, dopamine neurons release dopamine, providing a positive feedback signal to our brain, which increases the likelihood of related behaviors. In MARL, reward signals are analogous to the release of dopamine in the human brain (Dabney et al., 2020). When an agent takes action and obtains a positive outcome in the environment, it receives a reward signal. This reward signal can be a numerical value indicating the quality of the action, or it can be a more complex signal.

IMOE (Singh et al., 2004; Yang Z et al., 2022) is an RL exploration strategy inspired by the concept of intrinsic motivation in humans and animals. Its purpose is to encourage agents to explore the environment and learn new skills or acquire knowledge without relying on external rewards.In traditional RL, agents are driven to choose strategies by maximizing the external rewards provided by a predefined reward function. In IMOE, agents are driven by intrinsic curiosity or interest, which enables them to explore the environment and learn from it in the absence of external rewards. The connection between IMOE and RL algorithms such as temporal-difference (TD) (Sutton, 1984) lies in the balance between exploration and exploitation. TD algorithms maximize cumulative rewards based on predictions and estimates of future rewards, driving the learning process by updating value estimates through reward prediction errors. IMOE uses intrinsic motivation signals to guide exploration, encouraging agents to actively seek out novel or uncertain states and learn from them.

Compared to traditional noise-driven exploration methods (Mai et al., 2022; Wang SY et al., 2022), IMOE does not rely on external environmental rewards but uses the agent's intrinsic motivation and curiosity to drive exploration. The key idea is to define an intrinsic reward function that measures the novelty and uncertainty of the current state for the agent. The agent explores new states based on this intrinsic reward function with the goal of maximizing intrinsic rewards, rather than external environmental rewards. The Table 1 summarizes commonly used intrinsic motivation reward design methods.

Cooperative multiagent exploration (CMAE) (Liu IJ et al., 2021) does not rely on traditional noise-driven exploration but instead defines explicit shared target states, encouraging agents to coordinate their exploration efforts to achieve this goal. The target states are not randomly selected from the complete state space but are chosen from a predefined lower-dimensional constrained state space, gradually expanding the exploration range to higher-dimensional state spaces. Zhang HC et al. (2023) trained the Macro strategy controller (MaSC) network using heat maps based on successful experiences, while also using Manhattan distance as an intrinsic reward to encourage agents to guide micro strategies with macro strategies to achieve more complex cooperative behaviors.

Defining the intrinsic reward function is a key challenge faced by IMOE methods. The most direct method is to measure the novelty of the global observation information relative to historical observation states (Wang L et al., 2022). However, when the number of agents increases, the state space grows exponentially, making it very inefficient to find novel global observation information. If the approach is changed to measuring the novelty of local observation information during decentralized execution, it improves scalability. But due to partial observability, it cannot guide agents to cooperate. Multiagent with subgoals generated from experience replay (MASER) (Jeon et al., 2022) generates sub-goals within the experience replay buffer by both individual Q values of agents and the global Q value. A separate intrinsic reward function is generated for each sub-goal, guiding agents to explore optimal strategies to achieve their own subgoals, thereby maximizing the value of joint actions.

The Learnable intrinsic reward generation selec-

Unedited

Table 1 Categorization of intrinsic motivation exploration methods

Foundation	Algorithm type	Main principle	Representative algorithm
Counting	Density-based pseudo-count	A density model measures visitation, while a pseudo-count evaluates novelty. Agents receive exploration rewards based on pseudo-counts, encouraging visits to less-explored states.	Bellemare et al., 2016; Ostrovski et al., 2017
	Indirect pseudo-count	State novelty is gauged by the mismatch between observed states and the model's predictions, not by counting visits.	Fox L et al., 2018; Machado et al., 2020
	State abstraction	Agents generalize similar states, applying visitation counts to these abstractions rather than to the raw states.	Choi et al., 2019; Tang et al., 2017
Predictive model	Prediction error	High prediction error indicates a knowledge gap, driving further exploration.	Jiang H et al., 2022; Zhang HC et al., 2023
	Prediction outcome discrepancy	Agents use prediction discrepancies to guide exploration, with significant variances signaling areas of uncertainty for focused investigation.	Pathak et al., 2019; Ratzlaff et al., 2020
	Improvement of prediction accuracy	Agents adjust to outcome differences to enhance their models, reduce errors, and gradually increase precision.	Lopes et al., 2012; Graves et al., 2017
Information theory	Information gain	Agents pursue actions that maximize information gain, leveraging uncertainty to drive exploration and refine decision-making.	Zheng LL et al., 2021; Hairi et al., 2022
	Maximum entropy	Agents seek variety and high rewards, balancing exploration with exploitation to prevent settling on suboptimal strategies.	Hu et al., 2022; Liu BY et al., 2023; Xu P et al., 2023
	Mutual information	Agents leverage mutual information to inform actions that reveal more about the environment or peers, improving learning and collective decisions.	Eysenbach et al., 2019; Sharma et al., 2020

Unedited

tion (LIGS) framework (Mguni et al., 2022) includes an intrinsic reward generator that adopts a switching control mechanism. Based on the states and joint action history of all agents, it can quickly identify key states and assign intrinsic rewards to these states. The goal of the generator is to maximize the joint external return of the agents, which can guide the agents towards learning coordinated and optimal joint strategies. Hu et al. (2022) proposed a reward estimation for multiple action branches, modeling the reward distribution for all action branches of each agent to reduce the impact of reward uncertainty. Samples are taken from the reward distributions of different action branches according to the selection probabilities under the current policy. Then, a policy-weighted reward aggregation method is used to weight and aggregate the environmental rewards and the sampled rewards to obtain a stable training signal. Episodic MARL with curiosity-driven (EMC) (Zheng LL et al., 2021) uses the prediction error of the decomposed individual-agent Q values as an intrinsic reward to drive collaborative exploration and uses episodic memory to accelerate strategy learning. The intrinsic reward r_{int} of EMC is as follows:

$$r_{\text{int}} = \frac{1}{N}\sum_{i=1}^{N} ||\widetilde{Q}_i(\tau_i,\cdot) - Q_i^{\text{ext}}(\tau_i,\cdot)||^2, \qquad (17)$$

where $Q_i(\tau_i, a_i)$ represents the individual Q value of agent i, r_{int} represents the intrinsic reward based on the prediction error of the Q value, H represents the memory target provided by episodic memory, and θ represents the parameters of the inference module.

In multiagent tasks, the reward function usually depends only on a subspace of the state space as a structural prior knowledge. Subspace-aware multiagent exploration (SAME) (Xu P et al., 2023) proposed a new entropy exploration objective function $J_{wt}(\pi)$ that encourages exploration in subspaces

of the state space. This objective function assigns greater weight to subspaces with higher uncertainty, thus encouraging agents to explore these subspaces futher. However, direct optimization of $J_{wt}(\pi)$ is computationally challenging, so SAME provides an algorithm with lower computational complexity that promotes exploration by improving a lower bound of $J_{wt}(\pi)$. Since it only needs to consider one-dimensional subspaces for each dimension, rather than dealing with high-dimensional subspaces, the computational burden is reduced. Lazy agents avoidance through influencing external states (LAIES) (Liu BY et al., 2023) starts from the perspective of lazy agents and analyzes the problems in MARL under sparse reward settings, where the strategies learned by agents may not actively participate in team cooperation or strategies learned by them may contribute little to the overall system performance. LAIES divides the state into internal and external states and establishes a causal relationship graph between agents and the environment. It defines lazy agents and lazy teams mathematically. Two intrinsic reward incentive mechanisms are proposed: individual diligence intrinsic (IDI) and cooperative diligence intrinsic (CDI), which encourage both the individuals and the team as a whole to influence the external environment.

Ineffective exploration methods not only reduce the learning efficiency of agents but also prevent them from learning optimal strategies. In single-agent environments, this can be addressed by methods such as increasing the exploration rate (Bellemare et al., 2016; Haarnoja et al., 2018; Zheng LL et al., 2021; Hao JY et al., 2024), but these methods cannot be directly applied in MARL tasks. MAVEN (Mahajan et al., 2019) is an improvement over QMIX (Rashid et al., 2018), addressing the issue that QMIX is unable to explore effectively due to its monotonicity constraint. MAVEN introduces a hierarchical control of the latent space, whereby once the latent variable is determined, each joint action-value function can be considered as a joint exploration behavior pattern that persists throughout the entire episode. Furthermore, MAVEN learns a series of diverse behaviors by maximizing the mutual information between trajectories (sequences of observations and actions) and latent variables, allowing MAVEN to achieve exploration while adhering to constraints.

Chen E et al. (2022) argue that when agents seek intrinsic rewards and engage in unnecessary exploration, performance can decline even when sufficient task rewards are available. They propose a principled constrained policy optimization method that can automatically adjust the importance of intrinsic rewards, suppressing intrinsic rewards when exploration is not needed and increasing them when exploration is necessary. This allows for a balance between intrinsic and task-based rewards without the need for manual tuning.

While intrinsic motivation provides a powerful mechanism for encouraging agent exploration, it is often necessary to complement this with carefully designed extrinsic rewards. This brings us to the topic of reward function reshaping, which aims to modify the external reward signals to guide agents toward desired behaviors more effectively.

3.2 Reward reshaping guide signal reshaping

Reward reshaping involves adjusting the reward signals received by agents to encourage them to take actions that are more conducive to collective goals. By carefully designing the reward function, agents of various forms can be encouraged to adopt complementary behaviors, thus achieving a higher level of collaboration. For example, if the actions of an individual agent have a positive impact on other agents, that action can receive additional rewards. This approach helps to overcome the myopia and selfish behavior that individual agents may face, guiding them toward more coordinated and collectively optimized behavior patterns. Therefore, reward reshaping becomes a key means of harmonizing individual and collective interests to promote robust cooperation.Commonly used reward-reshaping methods are summarized in the Table 2.

Yi et al. (2022) enables agents to share rewards with their neighboring agents to encourage cooperation through collective actions towards achieving global goals. For each agent, a high-level strategy learns how to share rewards with neighbors to decompose the global goal, while a low-level strategy learns how to optimize the sub-goals decomposed by the high-level strategy. Xiao BC et al. (2022) used an attention mechanism to redistribute rewards to generate denser signals. During each training iteration, sampling is performed from the buffer, and the agent attention module is utilized to measure the importance of the strategy taken by each agent in the

unedited

Table 2 Summary of reward shaping methods

Method	Principle	Drawback	Algorithm
Potential-based reward shaping	Adds an additional reward in addition to the standard reinforcement learning reward signal, based on the difference of a potential function that scores states.	Difficult to design; potential for local optima; dependent on domain knowledge	Ng et al., 1999; Wiewiora, 2003; Badnava et al., 2023
Roadmap of potential-based reward shaping	Introduces the difference in state potential value as an additional reward to accelerate agent learning without changing the original optimal policy.	Difficult to design; potential for local optima; dependent on domain knowledge	Wiewiora et al., 2003; Devlin and Kudenko, 2012
From reward functions to dynamic potentials	Adjusts rewards using a variable potential function to adapt to changes in the environment and agent behavior for more effective learning.	Difficult to design; high computational complexity; difficult to adjust	Harutyunyan et al., 2015
Relative entropy inverse reinforcement learning	Infers a reward function that motivates agent behavior by minimizing the relative entropy between expert behavior and agent policy.	High computational complexity; low sample efficiency; potential for local optima	Suay et al., 2016; Wang YX et al., 2023
Reward shaping via metalearning	Infers a reward function that motivates agent behavior by minimizing the relative entropy between expert behavior and agent policy.	High training complexity; poor transfer performance; difficult to tune	Zou et al., 2019; Li K et al., 2021

Unedited

current state. Then, the temporal attention module is used to represent the importance of the current state throughout the entire episode. By combining agent attention and temporal attention, it is possible to model the relationships between agents and between time steps simultaneously, evaluating the quality of the actions taken by agents in the current state. Agent time attention (ATA) (She et al., 2022) redistributes sparse, delayed global team rewards to each agent and each time step through an additional loss. The redistributed dense reward signals can help agents learn better strategies. ATA uses a transformer encoder to integrate the state and action information of each agent and then uses a transformer decoder to redistribute the global rewards. By explicitly learning how to redistribute global rewards to individuals and time steps, it enhances the MAS's ability to handle complex planning challenges.

In complex multiagent collaborative environments, accurately assessing each agent's contribution and providing appropriate rewards are difficult. Attention individual intrinsic reward mixing (AIIR-MIX) (Li W et al., 2023) designed a nonlinear mixing network that dynamically combines intrinsic rewards and environmental rewards into a global reward for each agent. An attention mechanism is used to generate more refined and dynamic intrinsic rewards for each agent, to more accurately reflect each agent's contribution. AIRMN (Li W et al., 2024) has designed a new intrinsic reward network based on the attention mechanism, which combines intrinsic and extrinsic rewards in a nonlinear and dynamic way. It enables the total reward to adapt to changes in the environment, and it dynamically returns precise intrinsic rewards to each agent, thus better addressing the credit assignment problem.

Many existing MARL algorithms assume shared joint observations and actions among agents, which is often not viable in practice. Although the objective function introduces a hyperparameter $\gamma \in [0, 1)$ as a discount factor to account for the diminishing value of future rewards, it may not be suitable for tasks sensitive to long-term average rewards. Considering this aspect, Hairi et al. (2022) considered an average reward setting in MARL and, under the premise of partially observable environments, proposed a consensus-based batch sampling actor-critic algorithm. Agents update their policies by averaging the TD errors across the same batch of samples. ElSayed-Aly and Feng (2022) provided a flexible and automated framework for reward shaping based on logic, which can transform task specifications

into corresponding reward functions. They express task specifications using linear temporal logic (LTL) and synthesize limit-deterministic büchi automata (LDBA) to monitor the progress of the agents' tasks. When an observed label violates the LDBA specification, the automaton transitions to a trap state and issues a significant negative reward; when an observed label conforms to the specification, a positive reward is given. All agents receive the same reward based on the LDBA state to guide them in coordinating the completion of the task.

Reward function reshaping offers a way to refine individual agent incentives, but in cooperative multiagent settings, we must also consider how rewards are distributed across the team. This leads us to examine team reward allocation methods, which address the challenge of fairly and effectively distributing rewards among multiple agents working towards a common goal.

3.3 Team reward allocation credit assignment

In MARL tasks, researchers tend to adopt a framework of centralized training with decentralized execution (Sunehag et al., 2017; Kuba et al., 2022; Yu et al., 2022) to mitigate the environmental dynamics brought about by interactions among agents, thereby enhancing the stability and performance of MARL systems. Under this framework, a global critic network is usually constructed to evaluate the joint action value. However, the global critic network faces the credit assignment problem. In the global critic network, the critic receives the joint observations and joint actions of all agents as input and outputs a joint action value, representing the team-level reward. This structure lacks the ability to differentiate the contributions of individual agents to the overall team performance. In a multi-agent environment, each agent's role and contribution vary, but the global Critic network can only provide a composite action value assessment, lacking a precise evaluation of each agent's individual contributions. This design not only fails to achieve detailed credit assignment but also cannot provide accurate learning feedback to each agent.

To address this issue, recent research often utilizes the relationships between agents to decompose team-level rewards into individual agent rewards (Sunehag et al., 2017; Rashid et al., 2018; Yang YD et al., 2020a; Zhang TJ et al., 2020; Kim et al., 2022; Han et al., 2022; Wang TH et al., 2022). One approach is to introduce additional functions that marginalize the influence of individual agent behaviors (Foerster et al., 2018; Shen SQ et al., 2022; Zohar et al., 2022), an alternative is to use attention networks to reallocate individual rewards (Rashid et al., 2020; She et al., 2022; Xiao BC et al., 2022; Li W et al., 2023). Furthermore, to more finely address the credit assignment problem, some studies have turned to the use of hierarchical MARL methods (Feng et al., 2022; Jeon et al., 2022). In this approach, high-level policies are responsible for proposing abstract guiding principles, while low-level policies translate these principles into specific actions. This multilevel framework helps to allocate credit between different decision-making granularities, thereby enhancing learning efficiency and the quality of the policy. With such a layered strategy, the system can achieve more detailed and dynamic credit assignments in complex multiagent environments. These methods improve the accuracy and fairness of team-level reward distribution, thereby better reflecting the individual contributions of each agent. By breaking down team-level rewards into individual-level reward signals, agents can more accurately understand the impact of their actions on the overall system, enabling them to better adjust their strategies and improve collaborative capabilities.

The decomposition of value function (Sunehag et al., 2017; Rashid et al., 2018; Son et al., 2019) is a method used for credit assignment in MARL . In MASs, each agent must make decisions, but due to the interactions among agents, how to fairly distribute credit becomes an important issue. The decomposition of value function addresses this by breaking down the global value function into local value functions for each agent, thus enabling the distribution of credit. Each agent makes decisions based on its own local value function. In this way, each agent only needs to focus on its local value function, without having to consider the global value function, thereby reducing the complexity of the problem.The Table 3 summarizes some mainstream value decomposition methods used for solving the reliability assignment problem.

Counterfactual baselines (Foerster et al., 2018; Hou et al., 2023; Zhang NM et al., 2023; Qiao et al., 2024) comprise a method used for credit assignment

Unedited

Table 3 Comparison of value decomposition methods for credit assignment algorithms

Algorithm	Key technique	Drawback
VDN (Sunehag et al., 2017)	Summation of linearly decomposed agent values, parameter sharing among agents	Low efficiency, few game scenarios satisfy linear decomposition.
QMIX (Rashid et al., 2018)	Utilizes mixing and hypernetwork, IGM hypothesis, monotonicity constraints	The conditions for monotonicity constraints are too strong, may not converge to the optimal strategy.
WQMIX (Rashid et al., 2020)	Weighted mapping, feedforward network replaces mixing network, attention mechanism	High exploration requirements for the algorithm.
QTRAN (Son et al., 2019)	Self-attention mechanism, introduction of counterfactual methods	Fails to overcome the drawbacks of VDN and QMIX, high convergence conditions.
Qatten (Yang YD et al., 2020c)	Multihead fusion network, feedforward network based on global state	Lacks exploration capability.
MAVEN (Mahajan et al., 2019)	Multihead attention mechanism, variational exploration	Introduces a hierarchically controlled latent space, increasing algorithm complexity and training difficulty.
QPLEX (Wang JH et al., 2021b)	Monotonic value function decomposition, IGM based on advantage	Anchors the current Q_{tot} estimate as the upper bound of the entire function space, prone to local optima.
QPD (Yang YD et al., 2020b)	Path integral gradient decomposition, multichannel critic network	Using integral gradient methods reduces computational efficiency.
ROMA (Wang TH et al., 2020b)	Introduces roles, action clustering, role adaptation	Role division from the joint state-action space is extremely inefficient.
RODE (Wang TH et al., 2021)	Introduces roles, action clustering, two-level hierarchy	Role division and algorithm training cannot be end-to-end.
DAVE (Xu ZW et al., 2023)	Ignores IGM hypothesis, anti-self exploration	Sensitive to exploration coefficient settings.

unedited

in MARL . In MASs, due to the interactions and joint decision-making among agents, it is difficult to accurately evaluating the contribution of each agent to the overall system performance is difficult. Counterfactual baselines modify the strategy of each agent, changing it to an alternative strategy, and then determine the difference in outcomes between the actual strategy and the alternative strategy. Based on the difference in outcomes, the contribution of each agent is inferred, which allows for the fair distribution of rewards to agents.

4 Cooperative objective optimization methods

In MARL environments, a complex array of decision-making and learning processes are typically involved, wherein agents must collaborate to optimize a shared objective function. To effectively drive cooperative agent systems towards common goals, researchers have developed a variety of optimization strategies aimed at enhancing collaborative efficiency, strengthening strategy stability, and ensuring that the final behavioral policies achieve optimal synergy under varying environmental conditions and task requirements. In this section, we will delve into the latest advancements in cooperative objective optimization methods for MAS. Firstly, we introduce multiagent trust region optimization strategies, which are designed to balance exploration and exploitation, and prevent excessive optimization during the policy update process. Secondly, we highlight the optimization algorithms for communication among collaborative agents, which provide effective protocols and structures for information exchange between agents, facilitating both knowledge sharing and decision-making collaboration. Following this,

we discuss compatibility analysis and optimization methods for cooperative strategies, which focus on evaluating and adjusting the compatibility between agent policies to ensure the coordination and consistency of collective behavior.

4.1 Trust region optimization policy update stability

In MARL, each agent constantly learns and adjusts its strategy, influencing and interfering with one another, resulting in a continuously changing overall system environment. Since the actions and policy updates of an agent can change the environment for other agents, the environment that each agent faces is dynamic, making it difficult to maintain stability in the learning process, thus leading to the issue of nonstationarity.

Researchers use trust region methods (Schulman et al., 2015; Li WH et al., 2022) to ensure that agents gradually improve their performance during the learning process, rather than experiencing degradation or stagnation in performance. This enhances the stability and reliability of the learning process, allowing agents to progressively approach the optimal strategy or achieve a predetermined performance level. Through multiple iterations, agents can gradually optimize their policies and address the issue of nonstationarity.A summary of methods used for representing monotonic boundary constraints can be found in Table 4.

In distributed training and execution environments, independent proximal policy optimization (IPPO) (Witt et al., 2020) has shown impressive performance. Under this framework, each agent only needs to estimate its local value function. Despite some theoretical shortcomings (Rodriguez et al., 2023), IPPO's performance on the StarCraft Multi-Agent Challenge (SMAC) environment (Samvelyan et al., 2019) is be comparable to, or even better than, the best strategies. Its outstanding performance is partly attributed to its robustness to certain forms of environmental nonstationarity. However, distributed training and execution methods do not always converge. To address this issue, Nekoei et al. (Nekoei et al., 2023) proposed a sequentially updating learning method, where agents update their policies in sequence, ensuring that the learning process converges to an optimal solution for each agent. In this learning mode, when an agent updates its policy, all other agents' policies remain unchanged, thus reducing the nonstationarity issues that arise from simultaneous policy updates of other agents. In the ACE algorithm (Li CM et al., 2023), a serialized update strategy is adopted, simplifying the complex multiagent problem into a more manageable single-agent decision-making problem.

Yu et al. (2022) have provided evidence that trust region divergence possesses considerable promise for addressing nonstationarity in RL. The concept of trust region methods, which are designed to maintain stability in policy updates by restricting changes to a predefined "trust region," can be effectively adapted for MASs as demonstrated by the multiagent trust region policy optimization method (MATRPO) (Li HP and He, 2024), and related works (Roostaie and Ebadzadeh, 2021; Li WH et al., 2022). Yet, when juxtaposed with single-agent policy gradients, the estimators in multiagent contexts exhibit heightened variance (Yu et al., 2022). This challenge predominately stems from the obfuscation of an individual agent's reward signal due to the concurrent exploration behaviors of others within the environment. Such an interaction complicates the effective delineation of trust regions for each agent, thereby exacerbating the nonstationarity issue. Identification of strategies that can contain the trust region for each agent, despite the intertwined exploration activities, is crucial for the advancement of cooperative MARL.

Further advancements were made by Kuba et al. (2022), who introduced the dominance function decomposition theorem (Eq. (18)) and identified an optimal baseline for policy gradients (Eq. (19)), leading to a significant variance reduction in policy gradient estimates. Building upon these theoretical foundations, Building upon these theoretical developed a sequential rollout update mechanism. This approach allows agents scheduled later in the update sequence (later-order agents) to leverage the updated policies of those updated earlier (previous-order agents), thus integrating the cumulative knowledge across agent updates.

$$A^{\pi}_{i_{1:m}}(s, a_{i_{1:m}}) = \sum_{j=1}^{m} A^{\pi}_{i_j}(s, a_{i_{1:j-1}}, a_{i_j}). \tag{18}$$

unedited

Table 4 Summary of monotonic boundary constraint representation methods

Algorithm	Update	Sample efficiency	Monotonic bound
Wang XH et al., 2023	Sequential	Low	$4\varepsilon \sum_{i=1}^{n} \alpha_i \left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha_i)}\right)$
Yu et al., 2022	Simultaneous	High	$4\varepsilon \sum_{i=1}^{n} \frac{\alpha_i}{1-\gamma}$
Wu ZF et al., 2021	Simultaneous	High	$4\varepsilon \sum_{i=1}^{n} \alpha_i \left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\sum_{j=1}^{n} \alpha_j)}\right)$
Kuba et al., 2021	Sequential	High	$4\varepsilon \sum_{i=1}^{n} \alpha_i \left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\sum_{j=1}^{n} \alpha_j)}\right)$
Wang XH et al., 2023	Sequential	High	$4\varepsilon \sum_{i=1}^{n} \alpha_i \left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\sum_{j\in e_i \cup i} \alpha_j)}\right) + \sum_{i=1}^{n} \frac{\xi_i}{1-\gamma}$
Fu et al., 2022	Sequential	High	$4\varepsilon \sum_{i=1}^{n} \frac{\alpha_i}{1-\gamma}$
Zhuang et al., 2023	Simultaneous	High	$4\varepsilon \frac{\alpha_i}{1-\gamma}$
Yang TP et al., 2021	Simultaneous	High	$4\varepsilon \sum_{i=1}^{n} \frac{\alpha_i}{1-\gamma}$
Ye et al., 2023	Sequential	High	$4\varepsilon \sum_{i=1}^{n} \alpha_i \left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\sum_{j=1}^{n} \alpha_j)}\right)$

$$b_{\text{opt}}(s, a_i) = \frac{E_{a_i \sim \pi_i}[Q(s, a_i, a_{-i}) \nabla_{\theta_i} \log \pi_\theta(a_i|s)]}{E_{a_i \sim \pi_i}[\nabla_{\theta_i} \log \pi_\theta(a_i|s)]}. \tag{19}$$

This approach mandates that agents share parameters, which introduces constraints on the diversity of tasks that they are equipped to handle and poses difficulties in achieving convergence of cooperative strategies. The MAPPO (Yu et al., 2022) and IPPO (Witt et al., 2020) methodologies extend this paradigm, with all agents required to share parameters, which further limits the problem space that can be effectively addressed.

Kuba et al. (2021) provided a theoretical underpinning for the concerns associated with parameter sharing, demonstrating that it may lead to exponentially suboptimal outcomes. This is a pivotal finding, as it suggests that the common practice of parameter sharing in MASs could fundamentally compromise performance, leading to a degradation that scales with the complexity of the task and the number of agents involved.

In their seminal work, Kuba et al. (2021) enhanced the sequential policy update framework by leveraging the multiagent advantage function decomposition theorem. They established the monotonic improvement property of the joint policy and introduced the homogeneous-agent trust region policy optimization (HATRPO) and homogeneous-agent proximal policy optimization (HAPPO) algorithms (Kuba et al., 2021). These algorithms were shown to be capable of converging to a Nash equilibrium in cooperative multiagent settings.

While HAPPO offers guarantees for monotonic improvement on joint policy updates, it does not extend this assurance to individual agent policies. Addressing this limitation, agent by agent PPO (A2PO) (Wang XH et al., 2023) introduced a more rigorous monotonic improvement bound compared to HAPPO (as detailed in Table 4). This tighter constraint enables A2PO to theoretically enhance the efficiency of optimizing the collective strategy, providing a more robust solution to the nonstationary dilemmas inherent in MARL. Further contributing to the field, the targeted actor-distillation proximal policy optimization (TAD-PPO) framework (Ye et al., 2023) reconceptualizes the MDP into a structured single-agent MDP. This innovative perspective facilitates decentralized execution while preserving collaborative learning dynamics, allowing for the extraction and application of strategies from the con-

unedited

structed MDP to the multiagent environment.

While trust region methods provide a framework for stable policy updates in MASs, effective cooperation often requires explicit communication between agents. Let us now explore how communication mechanisms can be integrated into MARL algorithms to enhance coordination and collective performance.

4.2 Communication among collaborative agents

Effective communication is a cornerstone of interaction within MASs , bridging the gap between agents and their environment. It serves a dual purpose: fostering efficient information exchange and amplifying the collective efficacy of agent collaboration.

Expanding upon these ideas, Wang YT and Sartoretti (2022) proposed the utilization of the hidden states from multiple directional recurrent neural networks as a means of communication among agents. By using a multihop network topology, agents can receive and process information from all other agents at each timestep, significantly bolstering global collaborative efficacy. Moreover, there have been efforts to incorporate more sophisticated communication mechanisms into MARL. Jiang JC and Lu (2018) introduced an attention-based communication model, empowering agents to discern when to communicate and how to fuse shared information effectively. Perhaps most notably, Das et al. (2019) developed the targeted multiagent communication (TarMAC) model, enabling an agent to selectively determine its communication partners, thus streamlining the transmission of information within the MAS.

Contemporary role-based MARL approaches, such as those delineated by Wang et al. in RODE (Wang TH et al., 2021) and ROMA (Wang TH et al., 2020b), use predefined role structures to facilitate differentiated communication among roles. However, this predefinition can constrain the system's flexibility, particularly when confronted with varying team sizes or dynamic role changes. Addressing this challenge, Nguyen et al. (2022) advanced a novel framework that learns and transfers role assignments across teams of different scales. Initially, a role assignment network is cultivated within a small-scale team context; this network is subsequently transferred to a larger team setting, where it undergoes further refinement via environmental interactions.

Shao et al. (2022) introduced a self-organizing group mechanism, featuring a conductor election and message summary structure. In this system, leaders are elected every T times-teps to form temporary groups, each comprising a leader and followers. Within these groups, followers communicate with their leader, who then synthesizes and disseminates information, ensuring coordinated group actions. To mitigate the deleterious effects of poor-quality long-distance communication on decision-making, Xiao J et al. (2023) implemented a distance-based graph attention mechanism within the policy network. This module prioritizes agent proximity when establishing correlations, attenuating the influence of distant agents by focusing on the feature representations of neighbors, thereby enhancing the salience of pertinent information.

In scenarios wherein communication latency is unavoidable, the delay-aware communication model (DACOM) (Yuan TT et al., 2023) offers a solution. It operates within a delay-aware multiagent POMDP framework, accounting for both communication and action delays. The model's aggregation module uses an attention mechanism to ascertain the relevance of incoming messages, allowing for the prioritization and integration of the most critical information. Building upon the deep deterministic policy gradient(DDPG) framework, the work by Pesce and Montana (2020) enables explicit interagent communication. Pesce and Montana augmented DDPG with a memory module that facilitates the learning read-and- write operations during training. This innovation allows agents to collaboratively develop and utilize a shared environmental representation, streamlining collective decision-making processes. By integrating such a communication protocol, the utility of the DDPG algorithm is expanded to accommodate complex communication demands within MASs.

In MASs operating under communication constraints, agents often encounter challenges such as information delay or loss. Furthermore, not all situations necessitate global information for decision-making; local agent data are typically more pertinent. To optimize the decision-making process, attention mechanisms have emerged as a valuable research area, enabling agents to prioritize and thereafter process information selectively. These mechanisms, which include multihead attention, hierarchi-

Unedited

cal attention, and graph attention, empower agents to concentrate on the most task-relevant information amid the vast and intricate interactions.

Attention mechanisms have been pivotal in advancing MARL, with various studies affirming their efficacy. Transformer architectures, while powerful, often grapple with high space-time complexity (Das et al., 2019; Wen et al., 2022; Pu et al., 2023; Xiao J et al., 2023). To mitigate this, TransMix (Khan MJ et al., 2022) implements the additive self-attention mechanism proposed by Fastformer, which streamlines the complexity of the transformer encoder and enhances inference speed. TransMix, a network designed for joint action-value mixing, leverages this transformer-based approach to amalgamate individual agent value functions while capturing intricate environmental state dynamics through global-local context interactions. Mao et al. (2023) further refined the transformer's utility in MARL by bifurcating its structure into the inner and outer transformers. The inner transformer focuses on extracting spatial information from single observations at discrete timesteps, while the outer transformer processes sequential observation history to distill temporal information. This dual-structure design facilitates the extraction of comprehensive spatiotemporal representations, thereby enriching the decision-making process. Laskin et al. (2023) applied the transformer to reformulate offline RL as a sequence prediction challenge, significantly improving sample efficiency and training velocity. The model's strategic generative capabilities extend beyond replicating behavioral patterns from the dataset, demonstrating potential for innovating novel strategies distinct from the original training data. This capacity indicates that transformers can capture complex data patterns and extrapolate new solutions from these insights.

Tian et al. (2023) used the attention-based reward decomposition network with ensemble mechanism (ARDNEM) to parse the global reward into agent-specific components and reconstructs the joint trajectory with the decomposed prioritized experience replay (DPER). This enables agents to exchange and benefit from each other's optimal experiences. Pu et al. (2023) leverages the communication enhanced network (CEN) module and graph spatiotemporal long short-term memory network (GST-LSTM) to incorporate the states of neighboring agents, enriching the environmental information and stabilizing training. The CEN module, empowered by an enhanced attention mechanism, adeptly manages complex interactions among agents. Additionally, Du W et al. (2023) adopts a hierarchical graph attention structure, comprising agent-level and relationship-level attention modules. This dual mechanism discerns the significance of inter-agent relationships, accommodating different agent types. In TACO (Li DP et al., 2024), the attention framework is used to distill pertinent global information. Agents ascertain associated weights through self-attention and perform a weighted aggregation of other agents' hidden states to reconstruct global information. Over time, the model transitions from explicit communication to implicit cooperation, reducing reliance on direct information exchange during training.

The following Table 5 summarizes some of the key communication-based MARL methods.These communication-based MARL methods have shown promising results in improving agent coordination and system performance in various cooperative and competitive tasks. However, challenges remain in designing efficient communication protocols, handling the complexity of large-scale MASs, and addressing the trade-off between communication overhead and learning efficiency.

unedited

4.3 Cooperation compatibility—strategy alignment

Cooperative incompatibility represents a significant challenge in MARL research, stemming from potential strategic or behavioral conflicts among agents. These conflicts can arise from divergent goals, inconsistent action selection, or strategy misalignment (Lanctot et al., 2017; Zhao J et al., 2022). Traditional MARL training has often emphasized individual agent proficiency without sufficient focus on cooperative skill enhancement, leading to compatibility issues and hampering collective goal achievement. To counteract these issues, a plethora of strategies have been developed. Collaborative training and policy optimization techniques (Zeng et al., 2022; Li Y et al., 2023a; Zhang ZQ et al., 2023; Li Y et al., 2024) stand out as pivotal methods. These approaches aim to foster interagent cooperation and encourage actions that support collective task completion. However, even complete function families sat-

Table 5 Summary of communication-based MARL algorithms

Communication type	Algorithm name	Algorithm summary
Based on value function approximation	FedQMix (Cao et al., 2024)	Adds regularization penalties to punish the use of additional communication rounds, thereby improving the communication efficiency of agents.
	C2E (Du XQ et al., 2024)	Uses a set of critic networks that communicate with each other to estimate action values more accurately.
	DDRQN (Foerster et al., 2016)	Uses deep recurrent Q-networks to evaluate agent actions and communication strategies.
Based on policy gradient methods	DACO (Yuan TT et al., 2023)	Introduces the TimeNet component, which adjusts the waiting time for agents to receive messages from others, addressing delay-related uncertainties.
	BiCNet (Peng P et al., 2017)	Introduces bidirectional coordination networks to facilitate effective communication among multiple agents.
	MD-MADDPG (Pesce and Montana, 2020)	Uses shared memory as a communication channel, where agents read and provide information before action execution.
	Intrinsic A3C (Jaques et al., 2019)	Provides additional incentives for collaborative actions with high mutual information.
	MACC (Vanneste et al., 2020)	Uses counterfactual reasoning to train both action and communication strategies of agents.
Improve communication flexibility and learning efficiency	ATOC (Jiang JC and Lu, 2018)	Uses attention models to guide agent communication timing and information integration.
	MAC (Miuccio et al., 2024)	Agents learn to control communication protocols to communicate effectively while considering communication overhead.
	NASA (Abdel-Aziz et al., 2024)	Agents jointly learn adaptive communication protocols within a dynamic state space.
	RTS (Canese et al., 2024)	Optimizes communication protocols to reduce data loss between agents; exhibits strong robustness to the number of agents.
	I2C (Ding et al., 2020)	Proposes an independent inference communication mechanism, allowing agents to learn the behaviors of others without explicit communication.
Based on Attention	MACRL (Xiao J et al., 2023)	Employs graph attention mechanisms to generate agent aggregation vectors based on calculated inter-agent distance relevancy.
	AERL (Pu et al., 2023)	Uses spatiotemporal attention mechanisms to filter communication information and expand the communication range of agents.
	TarMAC (Das et al., 2019)	Allows agents to actively select the recipients of messages through a signature-based soft attention mechanism.

unedited

isfying the IGM principle may fall short of representing the optimal Q function for cooperative tasks (Fu et al., 2022). Alternatively, policy gradient methods augmented with auto-regressive policy modeling have the theoretical potential to represent any optimal policy. Yet, the reliance on sequential strategy dependencies may impede decentralized execution, limiting algorithmic scalability and practical application flexibility (Wang XH et al., 2023; Ye et al., 2023). Consequently, devising algorithms that balance scalability with the ability to effectively identify global optimum strategies in cooperative MARL environments remains a core research objective (Wang JH et al., 2021c).

Although current research has made progress on the problem of multiagent cooperation incompatibility, this issue remains a significant challenge in the field of MARL.Contemporary research primarily equips agents with the ability to learn self-interested strategies within a shared environment; however, ensuring compatibility and effective cooperation across diverse settings remains an area requiring further improvement. Approaches such as agent-level coordination MADDPG (ALC-MADDPG) (Zhang Y et al., 2021), deep implicit coordination graphs (DICG) (Li S et al., 2021), and cooperative MARL based on coordination degree (CMARL-CD) (Cui and Zhang, 2021) represent strides toward more-harmonious multiagent collaboration, concentrating on both global-level coordination and interagent local interactions. Specifically, ALC-MADDPG (Zhang Y et al., 2021) addresses the complexity of multiagent coordination by evaluating interagent correlations, facilitating strategic information exchange, and tailoring reward structures accordingly. DICG (Li S et al., 2021), on the other hand, uses a self-attention mechanism to compose a dynamic coordination graph. This graph's edges denote interaction intensities between agents, and subsequent application of graph convolutional networks on this framework aids in assimilating individual agent information and fortifying collective action. CMARL-CD (Cui and Zhang, 2021) eschews the arduous task of estimating a global Q-value function, instead empowering each agent to independently appraise the coordination level of its chosen actions. This assessment reflects the probability that an action will contribute optimally to the team's strategy, thus endorsing implicit cooperation without depending on extensive global state evaluation.

Addressing the influence of diverse behavioral styles within temporary cooperative teams, Fastap (Zhang ZQ et al., 2023) adopts an innovative training regimen. It clusters teammate strategies with varying behavioral patterns using the Chinese restaurant process (CRP), training controlled agents to identify and adapt to these patterns swiftly. By learning context encoders sensitive to teammate behavior, Fastap equips agents with the capability to modify their collaborative strategies accordingly, thereby mitigating issues of cooperation incompatibility.

To surmount the challenges of cooperative incompatibility, some studies have ventured into using pre-generated partner strategies. Notably, Charakorn et al. (2020) introduced the hierarchical multi-agent skill discovery (HMASD) algorithm, a two-tiered hierarchical framework. At the higher level, the HMASD algorithm utilizes a transformer structure to sequentially allocate skills, whereas the lower level is dedicated to uncovering valuable team and individual skills. This dual-level approach facilitates the coordination of disparate agent skills to accomplish complex multiagent tasks effectively.

Complementing this, Qu GN et al. (2020) advanced a scalable actor-critic algorithm tailored for expansive state-action spaces, commonly encountered in large-scale MASs. Ingeniously, the algorithm's computational complexity is confined to the scale of an agent's local neighborhood state-action space, rather than the vastness of the entire network's state-action space. This scalability is pivotal for the algorithm's viability in extensive multiagent scenarios. In their exploration of average reward objectives, Qu GN et al. (2020) demonstrate that, given limited agent interactions, the state-action function retains an exponentially decaying property. This characteristic underpins efficient learning and collaboration within intricate networks, asserting that agents can attain meaningful cooperation relying solely on local information. This insight holds significant implications for the pragmatic design of MASs in real-world contexts.

Furthermore, the quest to address cooperative incompatibility has also embraced the study of agents' action semantics. By leveraging representation learning methods, researchers aim to bolster robustness and scalability, thereby enhancing interagent cooperation (Wang WX et al., 2020; Hao XT

Unedited

et al., 2022). These methods seek to decipher the semantic information underlying actions, which could pave the way for more-nuanced and effective collaborative strategies in MARL environments.

Current investigations into achieving optimal coordination strategies in uncertain multiagent environments represent a vibrant area of inquiry. While existing MARL algorithms have made strides in facilitating agent coordination both globally and within local neighborhoods, the pursuit of granular, agent-to-agent coordination remains an open question. The intricacy of this problem is amplified by the multifaceted nature of interagent cooperation strategies, which intersect with domains such as game theory, mechanism design, and adaptive control.

5 Typical application scenarios in MARL

Collaborative MARL algorithms have garnered significant attention due to their applicability in various task scenarios. The ability of multiple agents to learn and adapt through interaction with the environment and each other has proven to be a powerful tool in solving complex problems. In this section, we explore specific applications of MARL in three distinct domains: smart games, smart cities, and national defense and military.

5.1 Smart games

In recent years, MARL has made significant strides in the development of smart games. Among the famous methods, DeepMind's DQN (Hessel et al., 2018) has been used to complete Atari games (Mnih et al., 2015), and AlphaStar (Arulkumaran et al., 2019) has competed with the world champion in Dota 2, showcasing the potential of RL in gaming.

MARL algorithms have demonstrated remarkable flexibility in adjusting strategies based on the lineup of agents in various StarCraft mission scenarios, achieving a 100% winning rate in most cases. Qu Y et al. (2024) released the real dataset of Honor of Kings and the offline RL benchmark HOKOFF, further advancing the field. Beyond MOBA games, MARL has also achieved excellent results in a wide range of board and sports games. Notable examples include Western chess (Perolat et al., 2022), Chinese chess (Li Y et al., 2023b), majong (Zhao XY and Holden, 2022), and poker (Zha et al., 2021). These achievements highlight the versatility and effectiveness of MARL in various gaming domains. However, these game scenarios often present significant challenges for MARL algorithms. For instance, the reward signal is typically only available at the end of the game, posing a severe challenge to real-time decision-making and strategy adjustment. Moreover, incomplete information and stochasticity further increase the complexity of reward modeling and policy learning.

In the realm of sports games, MARL has shown promising results in football (Zang et al., 2023; Chen JY et al., 2024; Jo et al., 2024) and basketball (Yeh et al., 2019). These environments present unique challenges due to their continuous action spaces and dynamic nature, making the design of effective reward signals particularly difficult. Furthermore, balancing short-term objectives (such as ball possession) with long-term goals (such as scoring) is crucial to avoid converging to suboptimal strategies.

o address these challenges, researchers have explored various approaches to enhance RL algorithms. For example, in the game Against the Cold, evolutionary RL combined with multiobjective optimization algorithms was used to create nonplayer characters (NPCs) with nonregularized behavior, enhancing the gaming experience (Zheng Y et al., 2019; Shen RM et al., 2020). Jeon et al. (2023) explored the problem of automated content balance in BOSS Raid games, demonstrating the potential of RL in game design and balance.

Recent advancements have pushed the boundaries of MARL even further. Park JS et al. (2023) created a virtual world consisting of 25 chat generative pre-trained transformer (ChatGPT) that completely simulated real human life, with agents carrying out their own activities and communicating with each other similar to humans. Zhu et al. (2023) imitated how humans solve complex problems in the real world, achieving higher learning efficiency and unlimited scalability. Their intelligent agents outperformed all previous agents in the game "Minecraft," showcasing the power of MARL in complex environments.

5.2 Smart cities and industry

MARL has found numerous applications in the development of smart cities and industries. Ren Y et al. (2024) investigated the robustness of RL in

unedited

multisignal control systems, highlighting its potential for enhancing the reliability and stability of urban infrastructure.

In the field of autonomous vehicles, Li Z et al. (2024) combined Q-learning and LSTM networks to enhance the driving decision-making capabilities of unmanned vehicles, enabling them to effectively adapt to dynamic and complex traffic conditions. This approach addresses the challenge of delayed rewards by leveraging the temporal memory capabilities of LSTM networks. Chen CQ et al. (2024) used RL algorithms to calculate competitive quotes for unmanned online ride-hailing vehicles in real time, based on current traffic flow and passenger flow, thereby optimizing transportation service efficiency.

Xiao BD et al. (2024) used MARL with value decomposition to construct a car network, considering the dynamic topological characteristics of the network during decision-making. By incorporating random graph neural networks, they effectively captured the underlying dynamics of network characteristics and improved the system's flexibility in response to environmental fluctuations. This approach provides a reference for extracting effective reward information from large-scale information flows.

In the domain of energy management, Yang NK et al. (2023) applied RL to implement a multienergy management strategy for hybrid electric vehicles, optimizing their energy efficiency and performance. The infrastructure management planning (IMP) environment (Leroy et al., 2024) provided a platform for MARL in large-scale infrastructure management plans, enabling the optimization of resource allocation and decision-making in complex urban systems.

RL has also found applications in robotics and industrial automation. Gu et al. (2023, 2024) improved the PPO algorithm to address robot control tasks, ensuring safe team behavior by meeting the safety constraints of each robot while maximizing team rewards. This approach utilizes trust region methods to balance safety constraints and credit allocation problems. Guo et al. (2023) solved the problem of multirobot patrolling using cooperative multiagent reinforcement learning and sequential decision-making. Cao et al. (2024) extended the functional control function to a MARL algorithm, enhancing the robot's safe movement and planning control capabilities in warehouse environments.

Despite the advantages of RL in industrial applications, challenges persist, such as high-dimensional parameter spaces and difficulties in agent convergence. To address these issues, Park S et al. (2024) proposed a quantum MARL algorithm that simplifies action dimensions by projecting value measurements and improves parameter utilization. This novel approach offers new perspectives on tackling the scalability and efficiency challenges in complex MARL scenarios.

As research in MARL progresses, its applications in smart cities and industries are expected to expand further. From optimizing urban infrastructure and transportation systems to enhancing energy management and industrial automation, RL holds immense potential for creating more efficient, sustainable, and intelligent urban environments and industrial processes.

5.3 National defense and military

In recent years, the application of MARL in the military domain has made significant progress, particularly in the scenario of drone air combat. Researchers have explored various approaches to enhance the effectiveness and cooperation of drones in different combat situations.

Wang BL et al. (2024) addressed the impact of combat cycles on air combat strategies by constructing an evolutionary RL algorithm. This algorithm enables drones to generate effective maneuver strategies based on the state changes of opponents, ultimately defeating them. Gong et al. (2023) combined the VDN algorithm with expert collaborative air combat experience to improve the level of cooperation between drones in various combat scenarios.

In the context of joint sea crossing and landing missions, Liu HY et al. (2024) used a large-scale multiagent evolutionary RL method. By dividing the learning stages into multiple phases according to the size of the agents, they achieved remarkable results in these complex military operations. Zhou et al. (2024) improved the MAPPO algorithm by incorporating the average value of the best formation and performing sampling, resulting in better formation strategies.

Kong et al. (2023) designed a hierarchical MARL strategy that encompasses sub-strategies and advanced strategies. By training agents through a self-game method, they demonstrated effective cooperative behavior in multiple scenarios. Wang ES

unedited

et al. (2024) focused on the UAV game confrontation problem in the noncooperative game model, providing insights into drone interactions in adversarial settings.

Medhi et al. (2023) investigated the use of MARL algorithms to reduce threats in byzantine attack scenarios, highlighting the potential of RL in enhancing the resilience of military systems.

As the field of MARL continues to evolve, its applications in military defense are expected to expand. From enhancing drone cooperation and maneuver strategies to optimizing joint military operations and decision-making processes, RL holds immense potential for revolutionizing the way military forces operate and adapt to complex and dynamic combat environments. However, it is crucial to consider the ethical implications and potential risks associated with the use of AI in military contexts, ensuring that its development and deployment align with principles of international law and human rights.

6 Simulation environment in a fully cooperative scenario

In the realm of MARL, simulation environments serve as indispensable tools. They offer a controlled yet dynamic virtual setting wherein agents can interact, learn, and synchronize their actions. Such environments are pivotal for researchers endeavoring to unravel the complexities of agent behavior, strategy formulation, and system performance. Through rigorous experimentation and training facilitated by these simulators, deeper insights into multiagent dynamics can be obtained.

To cater to the diverse research requirements posed by different MARL challenges, an array of specialized simulation environments has been designed, as detailed in Table 6. For instance, environments with partial observability such as Overcooked and starCraft multi-agent challenge (SMAC) (Samvelyan et al., 2019) are leveraged to assess the robustness of MARL algorithms. To evaluate algorithm scalability, environments that accommodate large numbers of agents, such as MAgent (Zheng LM et al., 2018) and MARLÖ (Perez-Liebana et al., 2019), have been developed. Safety-critical aspects are explored within specialized settings such as Safe MAMuJoCo, Safe MARobosuite (Gu et al., 2023).

Moreover, simulation environments tailored to specific practical applications have also been established, addressing domains such as drone swarms (Gao et al., 2019), autonomous driving (Krajzewicz, 2010; Sukhbaatar et al., 2016; Wang TH et al., 2020a), and robotic control (Peng B et al., 2021; Gu et al., 2023). These environments provide invaluable platforms for both the advancement and the rigorous evaluation of MARL algorithms.

7 Conclusion and future prospects

In this comprehensive review, we have systematically explored the multifaceted domain of MARL, with a particular focus on the intricacies of reward function construction and collaboration optimization in fully cooperative scenarios. The synthesis of literature presented herein underscores the critical role of well-designed reward mechanisms and the necessity for sophisticated cooperative strategies to empower agents within such systems.

To conclude this survey, in the following text, we discuss future perspectives and open problems for this direction.

7.1 Research on reward reshaping and guidance under fully cooperative games

In the domain of RL, reward signals are crucial-they act as direct feedback from the environment, informing the agent of the efficacy of its actions with respect to a given task. When engaging with MARL, the clarity and quality of these signals become even more critical, particularly when faced with suboptimal reward structures. Ambiguous, sparse, or inconsistent reward signals can significantly hinder the ability of agents to discern and adopt effective behavioral strategies.

This paper delves into the challenges associated with navigating reward signals in MARL, categorizing them into four principal groups for researchers' consideration

7.1.1 Sparse reward signals

Agents in multiagent environments often encounter sparse rewards, receiving feedback only at infrequent intervals. Such sparsity obfuscates the connection between specific actions and their reward outcomes, complicating the learning process.

unedited

Table 6 Introduction to typical multiagent testing environment

Environment name	Action space	Original learning mode	Reward	Observation
SMAC (Samvelyan et al., 2019)	Discrete	Cooperative	Mixed	Partial
MPE (Lowe et al., 2017)	Hybrid	Mixed	Dense	Full
MAMuJoCo (Witt et al., 2021)	Continuous	Cooperative	Dense	Partial
GRF (Kurach et al., 2020)	Discrete	Mixed	Sparse	Full
SISL (Gupta et al., 2017)	Hybrid	Cooperative	Dense	Full
LBF (Papoudakis et al., 2022)	Discrete	Mixed	Dense	Partial
RWARE (Papoudakis et al., 2022)	Discrete	Cooperative	Sparse	Partial
MAgent (Zheng LM et al., 2018)	Discrete	Mixed	Dense	Partial
Pommerman (Resnick et al., 2018)	Discrete	Mixed	Sparse	Full
MetaDrive (Li QY et al., 2023)	Continuous	Collaborative	Dense	Partial
MATE (Pan et al., 2022)	Hybrid	Mixed	Dense	Partial
GoBigger (Zhang M et al., 2023)	Continuous	Mixed	Dense	Mixed
Overcooked (Carroll et al., 2019)	Discrete	Cooperative	Dense	Full
MAPDN (Wang JH et al., 2021a)	Continuous	Cooperative	Dense	Partial
Hide Seek (Baker et al., 2020)	Multidiscrete	Mixed	Dense	Partial

unedited

7.2 Team-level reward signals

In team-based cooperative MARL systems, rewards are commonly allocated at the group level rather than to individual agents. This collective approach can obscure the contribution of an agent's singular actions to the overall team reward, making it difficult for agents to assess their individual impact.

7.2.1 Deceptive reward signals

In some MASs, reward signals may be intentionally deceptive. Such signals can incentivize agents to engage in behavior that yields higher immediate rewards through the deception of others, potentially at the cost of long-term collective success.

7.2.2 Delayed reward signals

The multiagent environment might also feature delayed rewards, whereby the consequences of an agent's actions are not immediately apparent. A sequence of actions might only yield feedback after a significant delay, complicating the attribution of outcomes to specific behaviors.

7.3 Multitask learning in open-ended world

Open-world environments are characterized by their dynamic and evolving nature, wherein agents encounter interdependent tasks that demand simultaneous learning and execution. The efficacy of MARL in such settings hinges on the creation of algorithms that can generalize across diverse tasks, adapt to novel scenarios, and transfer knowledge efficiently among tasks. Consider the case of Minecraft, an exemplar of an open-world environment with vast exploration spaces. Agents in Minecraft face a multitude of challenges, including imperfect information, low-quality reward signals, and exceedingly large state spaces. Moreover, the complex interrelationships among objects and the dependencies of tasks and projects underscore the necessity for precise multi-step reasoning and planning efficiency. Agents are required to retain focus on their initial goals while mitigating distractions, leading to an intricate state space and an ambiguous action space. The ability to foresee long-term consequences and

to formulate plans based on both current and potential future states is crucial (Mao et al., 2022; **?**; Zhu et al., 2023). The task of mastering long-term planning in open-world settings remains formidable. Agents must collaboratively negotiate to establish a unified plan, accommodating each other's capabilities, and dynamically adjust this plan as necessary. Such collaboration could involve pre-assigned roles or roles that are dynamically adapted during plan execution. Conflicts in planning must be addressed to maintain coherence, and agents can provide mutual oversight to rectify any planning errors.

Knowledge acquisition and utilization are pivotal for agents operating in open worlds. They must be capable of learning from the environment through self-supervision, accumulating experience, and enriching the built-in common sense of language models. The integration of external knowledge bases further augments this process. The representation of knowledge, selection of relevant information, and its integration into planning are critical. Knowledge can also fortify the reasoning abilities of language models. Although current planning approaches largely rely on language models' common-sense knowledge, there is an evident need to enhance the use of environment-specific knowledge. The open-world setting poses a rigorous test for MASs, wherein long-term planning, multistep reasoning, knowledge capture and application, and collaborative effort are paramount. Future research should pivot toward cultivating intelligent agents that can navigate these complexities autonomously, assimilating knowledge and engaging in effective communication and cooperation with fellow agents.

7.4 Value alignment and security robustness of agents

Value alignment constitutes a principal objective in the sphere of MARL research and applications, particularly from a safety standpoint (Burns et al., 2024; Ji et al., 2024). The quintessential challenge lies in ensuring that the objectives programmed into AI systems are congruent with human values and ethical norms. Misalignment between the goals of intelligent agents and human expectations can precipitate decisions that contravene human values, culminating in unforeseen or hazardous outcomes. Consequently, the scholarly consensus underscores the imperative for AI systems to achieve alignment with human values as a precondition for a secure AI-entwined future. Two principal risks are associated with the process of agent value alignment. The first pertains to 'reward hacking,' whereby agents may exploit system loopholes to maximize rewards via unforeseen strategies (Ma et al., 2024). Such strategies, while numerically successful, may diverge from human intentions, illustrating the necessity for more exacting reward definitions and considerations of potential misconfigurations during design. The second risk involves 'power-seeking' behaviors, whereby strategically aware agents could endeavor to augment their influence over the environment, irrespective of their initial objectives (Hua et al., 2024). This pursuit of control can lead to actions that compromise human interests or those of other agents, escalating systemic uncertainty and potential safety hazards.

To mitigate these risks, future research should prioritize the integration of human-centric values and ethical guidelines into algorithmic design, ensuring that agents not only strive to maximize collective rewards but also respect individual rights and adhere to societal norms. A promising avenue is to draw insights from the social sciences and behavioral economics to inform the structuration of agent reward mechanisms and decision-making processes (Xu YZ et al., 2024). For instance, incorporating principles from social choice theory could ensure that agent decisions reflect fairness (Li GH et al., 2023), while cooperative game theory could be leveraged to harmonize the pursuit of individual optimization with the maximization of collective rewards.

Unedited

Contributors

Tao YANG and Xinhao SHI were the primary contributors in writing the paper and conducting the literature review. Qinghan ZENG, Yulin YANG, and Hongzhe LIU provided critical revisions and contributed to the discussion of the content. Cheng XU, as the corresponding author, oversaw the project, provided the research framework, and managed the editorial process. All the authors read and approved the final paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Abdel-Aziz MK, Elbamby MS, Samarakoon S, et al., 2024. Cooperative multi-agent learning for navigation via structured state abstraction. *IEEE Trans Commun*, 72(6):3454-3462. https://doi.org/10.1109/TCOMM.2024.3365520

Andrew AM, 1999. Reinforcement learning: an introduction by Richard S. Sutton and Andrew G. Barto, Adaptive Computation and Machine Learning series, MIT Press (Bradford Book), Cambridge, Mass., 1998, xviii + 322 pp, ISBN 0-262-19398-1, (hardback, £31.95). *Robotica*, 17(2):229-235. https://doi.org/10.1017/S0263574799211174

Arulkumaran K, Cully A, Togelius J, 2019. AlphaStar: an evolutionary computation perspective. Proc Genetic and Evolutionary Computation Conf Companion, p.314-315. https://doi.org/10.1145/3319619.3321894

Badnava B, Esmaeili M, Mozayani N, et al., 2023. A new potential-based reward shaping for reinforcement learning agent. Proc IEEE 13th Annual Computing and Communication Workshop and Conf, p.1-6. https://doi.org/10.1109/CCWC57344.2023.10099211

Baker B, Kanitscheider I, Markov TM, et al., 2020. Emergent tool use from multi-agent autocurricula. Proc 8th Int Conf on Learning Representations.

Bellemare MG, Srinivasan S, Ostrovski G, et al., 2016. Unifying count-based exploration and intrinsic motivation. Proc 30th Conf on Neural Information Processing Systems, p.29.

Bernstein DS, Givan R, Immerman N, et al., 2002. The complexity of decentralized control of markov decision processes. *Math Oper Res*, 27(4):819-840. https://doi.org/10.1287/moor.27.4.819.297

Burns C, Izmailov P, Kirchner JH, et al., 2024. Weak-to-strong generalization: eliciting strong capabilities with weak supervision. Proc 41st Int Conf on Machine Learning.

Canese L, Cardarilli GC, Di Nunzio L, et al., 2024. Resilient multi-agent RL: introducing DQ-RTS for distributed environments with data loss. *Sci Rep*, 14(1):1994. https://doi.org/10.1038/s41598-023-48767-1

Cao HH, Xiong H, Zeng WF, et al., 2024. Safe reinforcement learning-based motion planning for functional mobile robots suffering uncontrollable mobile robots. *IEEE Trans Intell Transp Syst*, 25(5):4346-4363. https://doi.org/10.1109/TITS.2023.3330183

Cao SH, Zhang HQ, Wen T, et al., 2024. FedQMIX: communication-efficient federated learning via multi-agent reinforcement learning. *High-Confid Comput*, 4(2):100179. https://doi.org/10.1016/j.hcc.2023.100179

Carroll M, Shah R, Ho MK, et al., 2019. On the utility of learning about humans for human-AI coordination. Proc 33rd Conf on Neural Information Processing Systems, p.32.

Charakorn R, Manoonpong P, Dilokthanakul N, 2020. Investigating partner diversification methods in cooperative multi-agent deep reinforcement learning. Proc 27th Int Conf on Neural Information Processing, p.395-402. https://doi.org/10.1007/978-3-030-63823-8_46

Chen CQ, Yang HN, Zhai CJ, et al., 2024. Competitive pricing for ride-sourcing platforms with MARL. *Transp Res Part C Emerg Technol*, 165:104697. https://doi.org/10.1016/j.trc.2024.104697

Chen E, Hong ZW, Pajarinen J, et al., 2022. Redeeming intrinsic rewards via constrained optimization. Proc 36th Conf on Neural Information Processing Systems, p.4996-5008.

Chen HB, Ji WK, Xu LF, et al., 2023. Multi-agent consensus seeking via large language models. https://doi.org/10.48550/arXiv.2310.20151

Chen JY, Xu ZL, Li YF, et al., 2024. Accelerate multi-agent reinforcement learning in zero-sum games with subgame curriculum learning. Proc 38th AAAI Conf on Artificial Intelligence, p.11320-11328. https://doi.org/10.1609/aaai.v38i10.29011

Choi J, Guo YJ, Moczulski M, et al., 2019. Contingency-aware exploration in reinforcement learning. Proc 7th Int Conf on Learning Representations.

Cui HY, Zhang Z, 2021. A cooperative multi-agent reinforcement learning method based on coordination degree. *IEEE Access*, 9:123805-123814. https://doi.org/10.1109/ACCESS.2021.3110255

Dabney W, Kurth-Nelson Z, Uchida N, et al., 2020. A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792):671-675. https://doi.org/10.1038/s41586-019-1924-6

Das A, Gervet T, Romoff J, et al., 2019. TarMAC: targeted multi-agent communication. Proc 36th Int Conf on Machine Learning, p.1538-1546.

de Witt CS, Gupta T, Makoviichuk D, et al., 2020. Is independent learning all you need in the starcraft multi-agent challenge? https://doi.org/10.48550/arXiv.2011.09533

de Witt CS, Peng B, Kamienny PA, et al., 2021. Deep multi-agent reinforcement learning for decentralized continuous cooperative control. https://doi.org/10.48550/arXiv.2003.06709

Devlin S, Kudenko D, 2012. Dynamic potential-based reward shaping. Proc 11th Int Conf on Autonomous Agents and Multiagent Systems, p.433-440.

Ding ZL, Huang TJ, Lu ZQ, 2020. Learning individually inferred communication for multi-agent cooperation. Proc 34th Int Conf on Neural Information Processing Systems, p.22069-22079.

Du W, Ding SF, Zhang CL, et al., 2023. Multiagent reinforcement learning with heterogeneous graph attention network. *IEEE Trans Neur Netw Learn Syst*, 34(10):6851-6860. https://doi.org/10.1109/TNNLS.2022.3215774

Du XQ, Chen HC, Xing YH, et al., 2024. A contrastive-enhanced ensemble framework for efficient multi-agent reinforcement learning. *Exp Syst Appl*, 245:123158. https://doi.org/10.1016/j.eswa.2024.123158

ElSayed-Aly I, Feng L, 2022. Logic-based reward shaping for multi-agent reinforcement learning. https://doi.org/10.48550/arXiv.2206.08881

Unedited

Eysenbach B, Gupta A, Ibarz J, et al., 2019. Diversity is all you need: learning skills without a reward function. Proc 7th Int Conf on Learning Representations.

Feng L, Xie YX, Liu B, et al., 2022. Multi-level credit assignment for cooperative multi-agent reinforcement learning. *Appl Sci*, 12(14):6938. https://doi.org/10.3390/app12146938

Foerster J, Farquhar G, Afouras T, et al., 2018. Counterfactual multi-agent policy gradients. Proc 32nd AAAI Conf on Artificial Intelligence, p.2974-2982. https://doi.org/10.1609/aaai.v32i1.11794

Foerster JN, Assael YM, de Freitas N, et al., 2016. Learning to communicate to solve riddles with deep distributed recurrent q-networks. https://doi.org/10.48550/arXiv.1602.02672

Fox L, Choshen L, Loewenstein Y, 2018. DORA the explorer: directed outreaching reinforcement action-selection. Proc 6th Int Conf on Learning Representations.

Fu W, Yu C, Xu ZL, et al., 2022. Revisiting some common practices in cooperative multi-agent reinforcement learning. Proc 39th Int Conf on Machine Learning, p.6863-6877.

Gao F, Chen S, Li MQ, et al., 2019. MaCA: a multi-agent reinforcement learning platform for collective intelligence. Proc IEEE 10th Int Conf on Software Engineering and Service Science, p.108-111. https://doi.org/10.1109/ICSESS47205.2019.9040781

Gibbons R, 1992. A Primer in Game Theory. Pearson Academic, New York, USA.

Gong ZH, Xu Y, Luo DL, 2023. UAV cooperative air combat maneuvering confrontation based on multi-agent reinforcement learning. *Unmann Syst*, 11(3):273-286. https://doi.org/10.1142/S2301385023410029

Gou Y, Zhang T, Yang TT, et al., 2022. A deep MARL-based power-management strategy for improving the fair reuse of UWSNs. *IEEE Internet Things J*, 10(7):6507-6522. https://doi.org/10.1109/jiot.2022.3226953

Graves A, Bellemare MG, Menick J, et al., 2017. Automated curriculum learning for neural networks. Proc 34th Int Conf on Machine Learning, p.1311-1320.

Gu SD, Grudzien Kuba J, Chen YP, et al., 2023. Safe multi-agent reinforcement learning for multi-robot control. *Artif Intell*, 319:103905. https://doi.org/10.1016/j.artint.2023.103905

Gu SD, Huang DY, Wen MN, et al., 2024. Safe multiagent learning with soft constrained policy optimization in real robot control. *IEEE Trans Ind Inf*, 20(9):10706-10716. https://doi.org/10.1109/TII.2024.3391934

Guo LX, Pan HX, Duan XM, et al., 2023. Balancing efficiency and unpredictability in multi-robot patrolling: a MARL-based approach. Proc IEEE Int Conf on Robotics and Automation, p.3504-3509. https://doi.org/10.1109/ICRA48891.2023.10160923

Gupta JK, Egorov M, Kochenderfer M, 2017. Cooperative multi-agent control using deep reinforcement learning. Proc Int Conf on Autonomous Agents and Multiagent Systems, p.66-83. https://doi.org/10.1007/978-3-319-71682-4_5

Haarnoja T, Zhou A, Abbeel P, et al., 2018. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. Proc 35th Int Conf on Machine Learning, p.1861-1870.

Hairi, Liu J, Lu ST, 2022. Finite-time convergence and sample complexity of multi-agent actor-critic reinforcement learning with average reward. Proc 10th Int Conf on Learning Representations.

Han DG, Lu CX, Michalak T, et al., 2022. Multiagent model-based credit assignment for continuous control. Proc 21st Int Conf on Autonomous Agents and Multiagent Systems, p.571-579.

Hao JY, Yang TP, Tang HY, et al., 2024. Exploration in deep reinforcement learning: from single-agent to multiagent domain. *IEEE Trans Neur Netw Learn Syst*, 35(7):8762-8782. https://doi.org/10.1109/TNNLS.2023.3236361

Hao XT, Mao HY, Wang WX, et al., 2022. Breaking the curse of dimensionality in multiagent state space: a unified agent permutation framework. https://doi.org/10.48550/arXiv.2203.05285

Harsanyi JC, 1967. Games with incomplete information played by "Bayesian" players, I–III Part I. The basic model. *Manage Sci*, 14(3):159-182. https://doi.org/10.1287/mnsc.14.3.159

Harutyunyan A, Devlin S, Vrancx P, et al., 2015. Expressing arbitrary reward functions as potential-based advice. Proc 29th AAAI Conf on Artificial Intelligence, p.2652-2658. https://doi.org/10.1609/aaai.v29i1.9628

Hessel M, Modayil J, Van Hasselt H, et al., 2018. Rainbow: combining improvements in deep reinforcement learning. Proc 32nd AAAI Conf on Artificial Intelligence, p.3215-3222. https://doi.org/10.1609/aaai.v32i1.11796

Hou YK, Wei ZW, Liu SY, et al., 2023. Cross-regional task offloading with multi-agent reinforcement learning for hierarchical vehicular fog computing. Proc IEEE Symp on Computers and Communications, p.272-277. https://doi.org/10.1109/ISCC58397.2023.10217881

Hu JF, Sun YC, Chen HC, et al., 2022. Distributional reward estimation for effective multi-agent deep reinforcement learning. Proc 36th Conf on Neural Information Processing Systems, p.12619-12632.

Hua WY, Fan LZ, Li LY, et al., 2024. War and peace (WarAgent): large language model-based multi-agent simulation of world wars. https://doi.org/10.48550/arXiv.2311.17227

Huang JB, Tan QL, Qi RJ, et al., 2024. RELight: a random ensemble reinforcement learning based method for traffic light control. *Appl Intell*, 54(1):95-112. https://doi.org/10.1007/s10489-023-05197-w

Icarte RT, Klassen TQ, Valenzano R, et al., 2022. Reward machines: exploiting reward function structure in reinforcement learning. *J Artif Intell Res*, 73:173-208. https://doi.org/10.1613/jair.1.12440

Jaques N, Lazaridou A, Hughes E, et al., 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. Proc 36th Int Conf on Machine Learning, p.3040-3049.

Jeon HC, Baek IC, Bae CM, et al., 2023. RaidEnv: exploring new challenges in automated content balancing for boss raid games. *IEEE Trans Games*, in press. https://doi.org/10.1109/TG.2023.3335399

Jeon J, Kim W, Jung W, et al., 2022. MASER: multi-agent reinforcement learning with subgoals generated from experience replay buffer. Proc 39th Int Conf on Machine Learning, p.10041-10052.

Unedited

Ji JM, Qiu TY, Chen BY, et al., 2024. AI alignment: a comprehensive survey. https://doi.org/10.48550/arXiv.2310.19852

Jia LY, Cai CT, Wang XM, et al., 2023. Multi-intent autonomous decision-making for air combat with deep reinforcement learning. *Appl Intell*, 53(23):29076-29093. https://doi.org/10.1007/s10489-023-05058-6

Jiang H, Liu YT, Li SZ, et al., 2022. Diverse effective relationship exploration for cooperative multi-agent reinforcement learning. Proc 31st ACM Int Conf on Information & Knowledge Management, p.842-851. https://doi.org/10.1145/3511808.3557292

Jiang JC, Lu ZQ, 2018. Learning attentional communication for multi-agent cooperation. Proc 32nd Int Conf on Neural Information Processing Systems, p.7265-7275.

Jo Y, Lee S, Yeom J, et al., 2024. FoX: formation-aware exploration in multi-agent reinforcement learning. Proc 38th AAAI Conf on Artificial Intelligence, p.12985-12994. https://doi.org/10.1609/aaai.v38i12.29196

Khan MJ, Ahmed SH, Sukthankar G, 2022. Transformer-based value function decomposition for cooperative multi-agent reinforcement learning in StarCraft. Proc 18th AAAI Conf on Artificial Intelligence and Interactive Digital Entertainment, p.113-119. https://doi.org/10.1609/aiide.v18i1.21954

Khan R, Khan N, Ahmad T, 2023. Communication in multi-agent reinforcement learning: a survey. *Nucleus*, 60(2):175-185.

Kim SH, Van Stralen N, Chowdhary G, et al., 2022. Disentangling successor features for coordination in multi-agent reinforcement learning. Proc 21st Int Conf on Autonomous Agents and Multiagent Systems, p.751-760.

Kong WR, Zhou DY, Du YJ, et al., 2023. Hierarchical multi-agent reinforcement learning for multi-aircraft close-range air combat. *IET Control Theory Appl*, 17(13):1840-1862. https://doi.org/10.1049/cth2.12413

Krajzewicz D, 2010. Traffic simulation with SUMO-simulation of urban mobility. In: Barceló J (Ed.), Fundamentals of Traffic Simulation. Springer, New York, p.269-293. https://doi.org/10.1007/978-1-4419-6142-6_7

Kuba JG, Wen MN, Meng LH, et al., 2021. Settling the variance of multi-agent policy gradients. Proc 35th Conf on Neural Information Processing Systems, p.13458-13470.

Kuba JG, Chen RQ, Wen MN, et al., 2022. Trust region policy optimisation in multi-agent reinforcement learning. Proc 10th Int Conf on Learning Representations.

Kurach K, Raichuk A, Stańczyk P, et al., 2020. Google research football: a novel reinforcement learning environment. Proc 34th AAAI Conf on Artificial Intelligence, p.4501-4510. https://doi.org/10.1609/aaai.v34i04.5878

Lanctot M, Zambaldi V, Gruslys A, et al., 2017. A unified game-theoretic approach to multiagent reinforcement learning. Proc 31st Int Conf on Neural Information Processing Systems, p.4193-4206.

Laskin M, Wang LY, Oh J, et al., 2023. In-context reinforcement learning with algorithm distillation. Proc 11th Int Conf on Learning Representations.

Leroy P, Morato PG, Pisane J, et al., 2024. IMP-MARL: a suite of environments for large-scale infrastructure management planning via MARL. Proc 37th Int Conf on Neural Information Processing Systems, article 2329.

Li CM, Liu J, Zhang YM, et al., 2023. ACE: cooperative multi-agent Q-learning with bidirectional action-dependency. Proc 37th AAAI Conf on Artificial Intelligence, p.8536-8544. https://doi.org/10.1609/aaai.v37i7.26028

Li DP, Xu ZW, Zhang B, et al., 2024. From explicit communication to tacit cooperation: a novel paradigm for cooperative MARL. Proc 23rd Int Conf on Autonomous Agents and Multiagent Systems, p.2360-2362.

Li GH, Hammoud HAAK, Itani H, et al., 2023. CAMEL: communicative agents for "mind" exploration of large language model society. https://doi.org/10.48550/arXiv.2303.17760

Li HP, He HB, 2024. Multiagent trust region policy optimization. *IEEE Trans Neur Netw Learn Syst*, 35(9):12873-12887. https://doi.org/10.1109/TNNLS.2023.3265358

Li K, Gupta A, Reddy A, et al., 2021. MURAL: meta-learning uncertainty-aware rewards for outcome-driven reinforcement learning. Proc 38th Int Conf on Machine Learning, p.6346-6356.

Li QY, Peng ZH, Feng L, et al., 2023. MetaDrive: composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Trans Pattern Anal Mach Intell*, 45(3):3461-3475. https://doi.org/10.1109/TPAMI.2022.3190471

Li S, Gupta JK, Morales P, et al., 2021. Deep implicit coordination graphs for multi-agent reinforcement learning. Proc 20th Int Conf on Autonomous Agents and Multi-Agent Systems, p.764-772.

Li W, Liu WY, Shao ST, et al., 2023. AIIR-MIX: multi-agent reinforcement learning meets attention individual intrinsic reward mixing network. Proc 14th Asian Conf on Machine Learning, p.579-594.

Li W, Liu WY, Shao ST, et al., 2024. Attention-based intrinsic reward mixing network for credit assignment in multiagent reinforcement learning. *IEEE Trans Games*, 16(2):270-281. https://doi.org/10.1109/TG.2023.3263013

Li WH, Wang XF, Jin B, et al., 2022. Dealing with non-stationarity in MARL via trust-region decomposition. Proc 10th Int Conf on Learning Representations.

Li Y, Zhang S, Sun JC, et al., 2023a. Cooperative open-ended learning framework for zero-shot coordination. Proc 40th Int Conf on Machine Learning, p.20470-20484.

Li Y, Xiong K, Zhang YP, et al., 2023b. Jiangjun: mastering Xiangqi by tackling non-transitivity in two-player zero-sum games. *Trans Mach Learn Res*, 2023.

Li Y, Zhang S, Sun JC, et al., 2024. Tackling cooperative incompatibility for zero-shot human-AI coordination. *J Artif Intell Res*, 80:1139-1185. https://doi.org/10.1613/jair.1.15884

Li Z, Wang QC, Wang JB, et al., 2024. A flexible cooperative marl method for efficient passage of an emergency CAV in mixed traffic. *IEEE Trans Intell Transp Syst*, 25(8):8898-8912. https://doi.org/10.1109/TITS.2024.3411487

unedited

Liu BY, Pu ZQ, Pan Y, et al., 2023. Lazy agents: a new perspective on solving sparse reward problem in multi-agent reinforcement learning. Proc 40th Int Conf on Machine Learning, p.21937-21950.

Liu HY, Li ZH, Huang KH, et al., 2024. Evolutionary reinforcement learning algorithm for large-scale multi-agent cooperation and confrontation applications. *J Supercomput*, 80(2):2319-2346. https://doi.org/10.1007/s11227-023-05551-2

Liu IJ, Jain U, Yeh RA, et al., 2021. Cooperative exploration for multi-agent deep reinforcement learning. Proc 38th Int Conf on Machine Learning, p.6826-6836.

Lopes M, Lang T, Toussaint M, et al., 2012. Exploration in model-based reinforcement learning by empirically estimating learning progress. Proc 25th Int Conf on Neural Information Processing Systems, p.206-214.

Lowe R, Wu Y, Tamar A, et al., 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. Proc 31st Int Conf on Neural Information Processing Systems, p.6382-6393.

Ma YJ, Liang W, Wang GZ, et al., 2024. Eureka: human-level reward design via coding large language models. Proc 12th Int Conf on Learning Representations.

Machado MC, Bellemare MG, Bowling M, 2020. Count-based exploration with the successor representation. Proc 34th AAAI Conf on Artificial Intelligence, p.5125-5133. https://doi.org/10.1609/aaai.v34i04.5955

Mahajan A, Rashid T, Samvelyan M, et al., 2019. MAVEN: multi-agent variational exploration. Proc 33rd Int Conf on Neural Information Processing Systems, article 684.

Mai V, Mani K, Paull L, 2022. Sample efficient deep reinforcement learning via uncertainty estimation. Proc 10th Int Conf on Learning Representations.

Mao HY, Wang C, Hao XT, et al., 2022. SEIHAI: a sample-efficient hierarchical AI for the MineRL competition. Proc 3rd Int Conf on Distributed Artificial Intelligence, p.38-51. https://doi.org/10.1007/978-3-030-94662-3_3

Mao HY, Zhao R, Chen H, et al., 2023. Transformer in transformer as backbone for deep reinforcement learning. https://doi.org/10.48550/arXiv.2212.14538

Medhi JK, Liu R, Wang QL, et al., 2023. Robust multiagent reinforcement learning for UAV systems: Countering byzantine attacks. *Information*, 14(11):623. https://doi.org/10.3390/info14110623

Mguni DH, Jafferjee T, Wang JH, et al., 2022. LIGS: learnable intrinsic-reward generation selection for multi-agent learning. Proc 10th Int Conf on Learning Representations, p.2022.

Miuccio L, Riolo S, Samarakoon S, et al., 2024. On learning generalized wireless MAC communication protocols via a feasible multi-agent reinforcement learning framework. *IEEE Trans Mach Learn Commun Netw*, 2:298-317. https://doi.org/10.1109/TMLCN.2024.3368367

Mnih V, Kavukcuoglu K, Silver D, et al., 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529-533. https://doi.org/10.1038/nature14236

Nekoei H, Badrinaaraayanan A, Sinha A, et al., 2023. Dealing with non-stationarity in decentralized cooperative multi-agent deep reinforcement learning via multi-timescale learning. Proc 2nd Conf on Lifelong Learning Agents, p.376-398.

Ng AY, Harada D, Russell S, 1999. Policy invariance under reward transformations: Theory and application to reward shaping. Proc 16th Int Conf on Machine Learning, p.278-287.

Nguyen D, Nguyen P, Venkatesh S, et al., 2022. Learning to transfer role assignment across team sizes. Proc 21st Int Conf on Autonomous Agents and Multiagent Systems, p.963-971.

Nian XH, Li MM, Wang HB, et al., 2024. Large-scale UAV swarm confrontation based on hierarchical attention actor-critic algorithm. *Appl Intell*, 54(4):3279-3294. https://doi.org/10.1007/s10489-024-05293-5

Oroojlooy A, Hajinezhad D, 2023. A review of cooperative multi-agent deep reinforcement learning. *Appl Intell*, 53(11):13677-13722. https://doi.org/10.1007/s10489-022-04105-y

Ostrovski G, Bellemare MG, van den Oord A, et al., 2017. Count-based exploration with neural density models. Proc 34th Int Conf on Machine Learning, p.2721-2730.

Pan XH, Liu M, Zhong FW, et al., 2022. MATE: benchmarking multi-agent reinforcement learning in distributed target coverage control. Proc 36th Conf on Neural Information Processing Systems.

Papoudakis G, Christianos F, Rahman A, et al., 2019. Dealing with non-stationarity in multi-agent deep reinforcement learning. https://doi.org/10.48550/arXiv.1906.04737

Papoudakis G, Christianos F, Schäfer L, et al., 2022. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. Proc 35th Conf on Neural Information Processing Systems.

Park JS, O'Brien JC, Cai CJ, et al., 2023. Generative agents: interactive simulacra of human behavior. Proc 36th Annual ACM Symp on User Interface Software and Technology, article 2. https://doi.org/10.1145/3586183.3606763

Park S, Kim JP, Park C, et al., 2024. Quantum multi-agent reinforcement learning for autonomous mobility cooperation. *IEEE Commun Mag*, 62(6):106-112. https://doi.org/10.1109/MCOM.020.2300199

Pathak D, Gandhi D, Gupta A, 2019. Self-supervised exploration via disagreement. Proc 36th Int Conf on Machine Learning, p.5062-5071.

Peng B, Rashid T, Schroeder de Witt C, et al., 2021. FACMAC: factored multi-agent centralised policy gradients. Proc 35th Conf on Neural Information Processing Systems, p.12208-12221.

Peng P, Wen Y, Yang YD, et al., 2017. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. https://doi.org/10.48550/arXiv.1703.10069

Perez-Liebana D, Hofmann K, Mohanty SP, et al., 2019. The multi-agent reinforcement learning in malmÖ (MARLÖ) competition. https://doi.org/10.48550/arXiv.1901.08129

Perolat J, de Vylder B, Hennes D, et al., 2022. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990-996. https://doi.org/10.1126/science.add4679

Pesce E, Montana G, 2020. Improving coordination in small-scale multi-agent deep reinforcement learning through

Unedited

memory-driven communication. *Mach Learn*, 109(9-10):1727-1747. https://doi.org/10.1007/s10994-019-05864-5

Pesce E, Montana G, 2020. Improving coordination in small-scale multi-agent deep reinforcement learning through memory-driven communication. *Mach Learn*, 109(9-10):1727-1747. https://doi.org/10.1007/s10994-019-05864-5

Pu ZQ, Wang HM, Liu Z, et al., 2023. Attention enhanced reinforcement learning for multi agent cooperation. *IEEE Trans Neur Netw Learn Syst*, 34(11):8235-8249. https://doi.org/10.1109/TNNLS.2022.3146858

Qiao WC, Huang M, Gao ZM, et al., 2024. Distributed dynamic pricing of multiple perishable products using multi-agent reinforcement learning. *Exp Syst Appl*, 237:121252. ==https://doi.org/10.1016/j.eswa.2023.121252

Qu GN, Lin YH, Wierman A, et al., 2020. Scalable multi-agent reinforcement learning for networked systems with average reward. Proc 34th Int Conf on Neural Information Processing Systems, article 175.

Qu Y, Wang BY, Shao JZ, et al., 2024. Hokoff: real game dataset from honor of kings and its offline reinforcement learning benchmarks. Proc 37th Int Conf on Neural Information Processing Systems, article 974.

Rădulescu R, Mannion P, Roijers DM, et al., 2020. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Auton Agent Multi-Agent Syst*, 34(1):10. https://doi.org/10.1007/s10458-019-09433-x

Rashid T, Samvelyan M, de Witt CS, et al., 2018. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning. Proc 35th Int Conf on Machine Learning, p.4295-4304.

Rashid T, Farquhar G, Peng B, et al., 2020. Weighted QMIX: expanding monotonic value function factorisation for deep multi-agent reinforcement learning. Proc 34th Int Conf on Neural Information Processing Systems, article 855.

Ratzlaff N, Bai QX, Li FX, et al., 2020. Implicit generative modeling for efficient exploration. Proc 37th Int Conf on Machine Learning, article 740.

Ren FY, Dong W, Zhao XD, et al., 2024. Two-layer coordinated reinforcement learning for traffic signal control in traffic network. *Exp Syst Appl*, 235:121111. https://doi.org/10.1016/j.eswa.2023.121111

Ren Y, Zhang H, Du LK, et al., 2024. Stealthy black-box attack with dynamic threshold against marl-based traffic signal control system. *IEEE Trans Ind Inf*. https://doi.org/10.1109/TII.2024.3413356

Resnick C, Eldridge W, Ha D, et al., 2018. Pommerman: a multi-agent playground. Proc 14th AAAI Conf on Artificial Intelligence and Interactive Digital Entertainment.

Rodriguez J, Koutsopoulos HN, Wang SH, et al., 2023. Cooperative bus holding and stop-skipping: a deep reinforcement learning framework. *Transp Res Part C Emerg Technol*, 155:104308. https://doi.org/10.1016/j.trc.2023.104308

Roostaie S, Ebadzadeh MM, 2021. EnTRPO: trust region policy optimization method with entropy regularization. https://doi.org/10.48550/arXiv.2110.13373

Samvelyan M, Rashid T, de Witt CS, et al., 2019. The starCraft multi-agent challenge. Proc 18th Int Conf on Autonomous Agents and MultiAgent Systems, p.2186-2188 .

Schulman J, Levine S, Moritz P, et al., 2015. Trust region policy optimization. Proc 32nd Int Conf on Machine Learning, p.1889-1897.

Shao JZ, Lou ZQ, Zhang HC, et al., 2022. Self-organized group for cooperative multi-agent reinforcement learning. Proc 36th Int Conf on Neural Information Processing Systems, article 413.

Sharma A, Gu SX, Levine S, et al., 2020. Dynamics-aware unsupervised discovery of skills. Proc 8th Int Conf on Learning Representations.

She J, Gupta JK, Kochenderfer MJ, 2022. Agent-time attention for sparse rewards multi-agent reinforcement learning. Proc 21st Int Conf on Autonomous Agents and Multiagent Systems, p.1723-1725.

Shen RM, Zheng Y, Hao JY, et al., 2020. Generating behavior-diverse game AIs with evolutionary multi-objective deep reinforcement learning. Proc 29th Int Joint Conf on Artificial Intelligence, p.3371-3377. https://doi.org/10.24963/ijcai.2020/466

Shen SQ, Qiu MW, Liu J, et al., 2022. ResQ: a residual q function-based approach for multi-agent reinforcement learning value factorization. Proc 36th Int Conf on Neural Information Processing Systems, article 395.

Shou ZY, Di X, 2020. Reward design for driver repositioning using multi-agent reinforcement learning. *Transp Res Part C Emerg Technol*, 119:102738. https://doi.org/10.1016/j.trc.2020.102738

Singh S, Barto AG, Chentanez N, 2004. Intrinsically motivated reinforcement learning. Proc 17th Int Conf on Neural Information Processing Systems, p.1281-1288.

Singh S, Jaakkola T, Littman ML, et al., 2000. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine learning*, 38:287-308.

Son K, Kim D, Kang WJ, et al., 2019. QTRAN: learning to factorize with transformation for cooperative multi-agent reinforcement learning. Proc 36th Int Conf on Machine Learning, p.5887-5896.

Suay HB, Brys T, Taylor ME, et al., 2016. Learning from demonstration for shaping through inverse reinforcement learning. Proc Int Conf on Autonomous Agents & Multiagent Systems, p.429-437.

Sukhbaatar S, Szlam A, Fergus R, 2016. Learning multiagent communication with backpropagation. Proc 30th Int Conf on Neural Information Processing Systems, p.2252-2260.

Sunehag P, Lever G, Gruslys A, et al., 2017. Value-decomposition networks for cooperative multi-agent learning. https://doi.org/10.48550/arXiv.1706.05296

Sutton RS, 1988. Learning to predict by the methods of temporal differences. *Machine learning*, 3:9-44.

Sutton RS, 1984. Temporal Credit Assignment in Reinforcement Learning. University of Massachusetts Amherst, Massachusetts, USA.

Tang HR, Houthooft R, Foote D, et al., 2017. #Exploration: a study of count-based exploration for deep reinforcement learning. Proc 31st Int Conf on Neural Information Processing Systems, p.2750-2759.

unedited

Tian Q, Kuang K, Liu FR, et al., 2023. Learning from good trajectories in offline multi-agent reinforcement learning. Proc 37th AAAI Conf on Artificial Intelligence, p.11672-11680. https://doi.org/10.1609/aaai.v37i10.26379

Vanneste S, Vanneste A, Mets K, et al., 2020. Learning to communicate using counterfactual reasoning. https://doi.org/10.48550/arXiv.2006.07200

Wang BL, Gao XZ, Xie T, 2024. An evolutionary multi-agent reinforcement learning algorithm for multi-UAV air combat. *Knowl-Based Syst*, 299:112000. https://doi.org/10.1016/j.knosys.2024.112000

Wang ES, Liu F, Hong C, et al., 2024. MADRL-based UAV swarm non-cooperative game under incomplete information. *Chin J Aeronaut*, 37(6):293-306. https://doi.org/10.1016/j.cja.2024.03.030

Wang JH, Xu WK, Gu YJ, et al., 2021a. Multi-agent reinforcement learning for active voltage control on power distribution networks. Proc 35th Int Conf on Neural Information Processing Systems, article 250.

Wang JH, Ren ZZ, Liu T, et al., 2021b. QPLEX: duplex dueling multi-agent Q-learning. Proc 9th Int Conf on Learning Representations.

Wang JH, Ren ZZ, Han BN, et al., 2021c. Towards understanding cooperative multi-agent q-learning with value factorization. Proc 35th Conf on Neural Information Processing Systems, p.29142-29155.

Wang JR, Hong YT, Wang JL, et al., 2022. Cooperative and competitive multi-agent systems: From optimization to games. *IEEE/CAA J Autom Sin*, 9(5):763-783. https://doi.org/10.1109/JAS.2022.105506

Wang L, Zhang YP, Hu YJ, et al., 2022. Individual reward assisted multi-agent reinforcement learning. Proc 39th Int Conf on Machine Learning, p.23417-23432.

Wang SY, Chen WY, Hu J, et al., 2022. Noise-regularized advantage value for multi-agent reinforcement learning. *Mathematics*, 10(15):2728. https://doi.org/10.3390/math10152728

Wang TH, Wang JH, Zheng CY, et al., 2020a. Learning nearly decomposable value functions via communication minimization. Proc 8th Int Conf on Learning Representations.

Wang TH, Dong H, Lesser V, et al., 2020b. ROMA: multi-agent reinforcement learning with emergent roles. Proc 37th Int Conf on Machine Learning, p.9876-9886.

Wang TH, Gupta T, Mahajan A, et al., 2021. RODE: learning roles to decompose multi-agent tasks. Proc 9th Int Conf on Learning Representations.

Wang TH, Zeng L, Dong WJ, et al., 2022. Context-aware sparse deep coordination graphs. Proc 10th Int Conf on Learning Representations.

Wang WX, Yang TP, Liu Y, et al., 2020. Action semantics network: considering the effects of actions in multiagent systems. Proc 8th Int Conf on Learning Representations.

Wang XH, Tian Z, Wan ZY, et al., 2023. Order matters: agent-by-agent policy optimization. Proc 11th Int Conf on Learning Representations.

Wang YT, Sartoretti G, 2022. FCMNet: full communication memory net for team-level cooperation in multi-agent systems. Proc 21st Int Conf on Autonomous Agents and Multiagent Systems, p.1355-1363.

Wang YX, Zeng ZX, Zhao QJ, 2023. Evaluating the perceived safety of urban city via maximum entropy deep inverse reinforcement learning. Proc 14th Asian Conf on Machine Learning, p.1085-1100.

Wen MN, Kuba J, Lin RJ, et al., 2022. Multi-agent reinforcement learning is a sequence modeling problem. Proc 36th Int Conf on Neural Information Processing Systems, article 1201.

White III CC, White DJ, 1989. Markov decision processes. *Eur J Oper Res*, 39(1):1-16. https://doi.org/10.1016/0377-2217(89)90348-2

Wiewiora E, 2003. Potential-based shaping and Q-value initialization are equivalent. *J Artif Intell Res*, 19:205-208. https://doi.org/10.1613/jair.1190

Wiewiora E, Cottrell G, Elkan C, 2003. Principled methods for advising reinforcement learning agents. Proc 20th Int Conf on Machine Learning, p.792-799.

Wu T, Zhou P, Liu K, et al., 2020. Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks. *IEEE Trans Veh Technol*, 69(8):8243-8256. https://doi.org/10.1109/TVT.2020.2997896

Wu ZF, Yu C, Ye DC, et al., 2021. Coordinated proximal policy optimization. Proc 35th Conf on Neural Information Processing Systems, p.26437-26448.

Xiao BC, Ramasubramanian B, Poovendran R, 2022. Agent-temporal attention for reward redistribution in episodic multi-agent reinforcement learning. Proc 21st Int Conf on Autonomous Agents and Multiagent Systems, p.1391-1399.

Xiao BD, Li RP, Wang F, et al., 2024. Stochastic graph neural network-based value decomposition for MARL in internet of vehicles. *IEEE Trans Veh Technol*, 73(2):1582-1596. https://doi.org/10.1109/TVT.2023.3312574

Xiao J, Yuan GH, He JH, et al., 2023. Graph attention mechanism based reinforcement learning for multi-agent flocking control in communication-restricted environment. *Inf Sci*, 620:142-157. https://doi.org/10.1016/j.ins.2022.11.059

Xu P, Zhang JG, Yin QY, et al., 2023. Subspace-aware exploration for sparse-reward multi-agent tasks. Proc 37th AAAI Conf on Artificial Intelligence, p.11717-11725. https://doi.org/10.1609/aaai.v37i10.26384

Xu X, Jia Y, Xu Y, et al., 2020. A multi-agent reinforcement learning-based data-driven method for home energy management. *IEEE Transactions on Smart Grid*, 11(4):3201-3211. https://doi.org/10.1109/TSG.2020.2971427

Xu YZ, Wang S, Li P, et al., 2024. Exploring large language models for communication games: an empirical study on werewolf. https://doi.org/10.48550/arXiv.2309.04658

Xu ZW, Zhang B, LI DP, et al., 2023. Dual self-awareness value decomposition framework without individual global max for cooperative MARL. Proc 37th Conf on Neural Information Processing Systems, p.36.

Yang NK, Han LJ, Liu R, et al., 2023. Multiobjective intelligent energy management for hybrid electric vehicles based on multiagent reinforcement learning. *IEEE Trans Transp Electrif*, 9(3):4294-4305. https://doi.org/10.1109/TTE.2023.3236324

Yang TP, Wang WX, Tang HY, et al., 2021. An efficient transfer learning framework for multiagent reinforcement learning. Proc 35th Int Conf on Neural Information Processing Systems, article 1302.

Unedited

Yang YD, Wen Y, Chen LH, et al., 2020a. Multi-agent determinantal Q-learning. Proc 37th Int Conf on Machine Learning, article 997.

Yang YD, Hao JY, Chen GY, et al., 2020b. Q-value path decomposition for deep multiagent reinforcement learning. Proc 37th Int Conf on Machine Learning, article 992.

Yang YD, Hao JY, Liao B, et al., 2020c. Qatten: a general framework for cooperative multiagent reinforcement learning. https://doi.org/10.48550/arXiv.2002.03939

Yang Z, Moerland TM, Preuss M, et al., 2022. When to go, and when to explore: the benefit of post-exploration in intrinsic motivation. https://doi.org/10.48550/arXiv.2203.16311

Ye JN, Li CH, Wang JH, et al., 2023. Towards global optimality in cooperative MARL with the transformation and distillation framework. https://doi.org/10.48550/arXiv.2207.11143

Yeh RA, Schwing AG, Huang J, et al., 2019. Diverse generation for multi-agent sports games. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4605-4614. https://doi.org/10.1109/CVPR.2019.00474

Yi YX, Li G, Wang YW, et al., 2022. Learning to share in multi-agent reinforcement learning. Proc 36th Int Conf on Neural Information Processing Systems, article 1100.

Yu C, Velu A, Vinitsky E, et al., 2022. The surprising effectiveness of PPO in cooperative multi-agent games. Proc 36th Int Conf on Neural Information Processing Systems, article 1787.

Yuan TT, Chung HM, Yuan J, et al., 2023. DACOM: learning delay-aware communication for multi-agent reinforcement learning. Proc 37th AAAI Conf on Artificial Intelligence, p.11763-11771. https://doi.org/10.1609/aaai.v37i10.26389

Yuan WL, Chen JX, Chen SF, et al., 2024. Transformer in reinforcement learning for decision-making: a survey. *Front Inf Technol Electron Eng*, 25(6):763-790. https://doi.org/10.1631/FITEE.2300548

Zang YF, He JM, Li K, et al., 2023. Automatic grouping for efficient cooperative multi-agent reinforcement learning. Proc 37th Conf on Neural Information Processing Systems, p.36.

Zeng SL, Chen TY, Garcia A, et al., 2022. Learning to coordinate in multi-agent systems: a coordinated actor-critic algorithm and finite-time guarantees. Proc 4th Annual Learning for Dynamics and Control Conf, p.278-290.

Zha DC, Xie JR, Ma WY, et al., 2021. DouZero: mastering DouDizhu with self-play deep reinforcement learning. Proc 38th Int Conf on Machine Learning, p.12333-12344.

Zhang HC, Li GZ, Liu CH, et al., 2023. HiMacMic: hierarchical multi-agent deep reinforcement learning with dynamic asynchronous macro strategy. Proc 29th ACM SIGKDD Conf on Knowledge Discovery and Data Mining, p.3239-3248.

Zhang KQ, Yang ZR, Başar T, 2021. Decentralized multi-agent reinforcement learning with networked agents: recent advances. *Front Inf Technol Electron Eng*, 22(6):802-814. https://doi.org/10.1631/FITEE.1900661

Zhang M, Zhang SH, Yang ZJ, et al., 2023. GoBigger: a scalable platform for cooperative-competitive multi-agent interactive simulation. Proc 11th Int Conf on Learning Representations.

Zhang NM, Shen YL, Du Y, et al., 2023. Counterfactual-attention multi-agent reinforcement learning for joint condition-based maintenance and production scheduling. *J Manuf Syst*, 71:70-81. https://doi.org/10.1016/j.jmsy.2023.08.011

Zhang TJ, Xu HZ, Wang XL, et al., 2020. Multi-agent collaboration via reward attribution decomposition. https://doi.org/10.48550/arXiv.2010.08531

Zhang Y, Yang QY, An D, et al., 2021. Coordination between individual agents in multi-agent reinforcement learning. Proc 35th AAAI Conf on Artificial Intelligence, p.11387-11394. https://doi.org/10.1609/aaai.v35i13.17357

Zhang ZQ, Yuan L, Li LH, et al., 2023. Fast teammate adaptation in the presence of sudden policy change. Proc 39th Conf on Uncertainty in Artificial Intelligence, p.2465-2476.

Zhao J, Zhao YP, Wang WX, et al., 2022. Coach-assisted multi-agent reinforcement learning framework for unexpected crashed agents. *Front Inf Technol Electron Eng*, 23(7):1032-1042. https://doi.org/10.1631/FITEE.2100594

Zhao LY, Chang TQ, Chu KX, et al., 2023. Survey of fully cooperative multi-agent deep reinforcement learning. *Comput Eng Appl*, 59(12):14-27 (in Chinese). https://doi.org/10.3778/j.issn.1002-8331.2209-0186

Zhao XY, Holden SB, 2022. Towards a competitive 3-player mahjong AI using deep reinforcement learning. Proc IEEE Conf on Games, p.524-527. https://doi.org/10.1109/CoG51982.2022.9893576

Zheng LL, Chen JR, Wang JH, et al., 2021. Episodic multi-agent reinforcement learning with curiosity-driven exploration. Proc 35th Int Conf on Neural Information Processing Systems, article 287.

Zheng LM, Yang JC, Cai H, et al., 2018. MAgent: a many-agent reinforcement learning platform for artificial collective intelligence. Proc 32nd AAAI Conf on Artificial Intelligence, p.8222-8223. https://doi.org/10.1609/aaai.v32i1.11371

Zheng Y, Xie XF, Su T, et al., 2019. Wuji: automatic online combat game testing using evolutionary deep reinforcement learning. Proc 34th IEEE/ACM Int Conf on Automated Software Engineering, p.772-784. https://doi.org/10.1109/ase.2019.00077

Zhou YM, Yang F, Zhang CY, et al., 2024. Cooperative decision-making algorithm with efficient convergence for UCAV formation in beyond-visual-range air combat based on multi-agent reinforcement learning. *Chin J Aeronaut*, 37(8):311-328. https://doi.org/10.1016/j.cja.2024.04.008

Zhu XZ, Chen YT, Tian H, et al., 2023. Ghost in the minecraft: generally capable agents for open-world environments via large language models with text-based knowledge and memory. https://doi.org/10.48550/arXiv.2305.17144

Zhuang ZF, Lei K, Liu JX, et al., 2023. Behavior proximal policy optimization. Proc 11th Int Conf on Learning Representations.

Under Review

Zohar R, Mannor S, Tennenholtz G, 2022. Locality matters: a scalable value decomposition approach for cooperative multi-agent reinforcement learning. Proc 36th AAAI Conf on Artificial Intelligence, p.9278-9285. https://doi.org/10.1609/aaai.v36i8.20915

Zou HS, Ren TZ, Yan D, et al., 2019. Reward shaping via meta-learning. https://doi.org/10.48550/arXiv.1901.09330

unedited