多智能体强化学习:从基础理论到前沿算法

韩光洁[1],朱胜超[2],林 川[3],江金芳[1]
(1. 河海大学信息科学与工程学院,江苏常州 213200;2. 河海大学计算机与软件学院,江苏南京 211100;
3. 东北大学软件学院,辽宁沈阳 110169)

摘 要: 多智能体强化学习(Multi-Agent Reinforcement Learning,MARL)作为处理复杂动态环境中智能体协作与竞争问题的重要框架,近年来在理论与应用上取得快速发展,并在自动驾驶、群体机器人、智能调度与对抗博弈等领域展现出广阔前景. 然而,多智能体系统中普遍存在环境非平稳、策略强耦合、信用分配困难和安全约束复杂等问题,使得MARL相较于单智能体强化学习面临更大挑战. 本文首先梳理了MARL的基础建模与理论框架,从马尔可夫博弈、部分可观测马尔可夫博弈等形式化描述出发,结合集中式训练、分布式执行和基于通信的协同决策等典型范式,对现有方法在信息利用、计算复杂度与收敛性质等方面进行对比分析,并围绕价值分解、策略梯度、多智能体信用分配和通信建模等核心技术进行归纳. 在此基础上,本文重点总结了若干前沿研究方向:一是基于大语言模型(Large Language Model,LLM)的MARL,通过引入LLM的知识推理和高层规划能力,用于任务分解、策略引导及自然语言通信,以提升智能体在开放环境中的泛化与协作能力;二是基于元学习的MARL,面向多任务与分布迁移场景,关注策略对新任务、新队友或新对手的快速适应,通过学习“会学习的初始化”或适应规则提高样本效率;三是基于可解释性的MARL,利用注意力可视化、因果分析和规则抽取等方法增强决策过程透明度,为策略审计、人机协同与安全监管提供支持;四是大规模MARL的应用与部署,聚焦智能体数量和状态维度急剧增长带来的训练效率、通信开销与可扩展性问题,探索分层结构、群体建模和并行训练等机制;五是多智能体安全强化学习,从约束满足、风险控制和稳健性出发,研究在对抗扰动、不确定性和策略博弈下的安全决策. 最后,本文结合协作与竞争两类典型应用场景,讨论了MARL在真实系统落地中面临的样本效率不足、仿真到现实迁移困难、公平性与稳态博弈分析不足等挑战,旨在为后续MARL的理论研究与工程应用提供系统参考.

关键词: 多智能体强化学习;马尔可夫博弈;大语言模型;元学习;可解释性;多智能体安全强化学习
基金项目: 国家自然科学基金(No.U22A2011)
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112(2025)12-4756-31
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250418

Multi-Agent Reinforcement Learning: From Foundational Theory to Cutting-Edge Algorithms

HAN Guang-jie[1], ZHU Sheng-chao[2], LIN Chuan[3], JIANG Jin-fang[1]
(1. *College of Information Science and Engineering*, *Hohai University*, *Changzhou*, *Jiangsu* 213200, *China*;
2. *College of Computer Science and Software Engineering*, *Hohai University*, *Nanjing*, *Jiangsu* 211100, *China*;
3. *Software College*, *Northeastern University*, *Shenyang*, *Liaoning* 110169, *China*)

Abstract: Multi-Agent Reinforcement Learning (MARL), as an important framework for handling the problems of agent cooperation and competition in complex dynamic environments, has achieved rapid development in both theory and application in recent years, and has shown broad prospects in fields such as autonomous driving, swarm robotics, intelligent scheduling, and adversarial games. However, problems such as environmental non-stationarity, strong policy coupling, difficult credit assignment, and complex safety constraints are widespread in multi-agent systems, making MARL face greater challenges compared to single-agent reinforcement learning. This paper first combs through the foundational modeling and theoretical framework of MARL, starting from formal descriptions such as Markov games and partially observable Markov games, and combining typical paradigms such as centralized training with decentralized execution, and communication-based cooperative decision-making, to conduct a comparative analysis of existing methods in terms of information utiliza-

收稿日期:2025-05-26;录用日期:2025-12-05;责任编辑:覃怀银

tion, computational complexity, and convergence properties, and summarizes the core technologies such as value decomposition, policy gradients, multi-agent credit assignment, and communication modeling. On this basis, this paper focuses on summarizing several frontier research directions. The first is Large Language Models (LLMs)-based MARL, which, by introducing the knowledge reasoning and high-level planning capabilities of LLMs, is used for task decomposition, policy guidance, and natural language communication, to enhance the generalization and collaboration capabilities of agents in open environments. The second is MARL based on meta-learning, facing multi-task and distribution shift scenarios, focusing on the rapid adaptation of policies to new tasks, new teammates, or new opponents, improving sample efficiency by learning "learn-to-learn" initializations or adaptation rules. The third is MARL based on explainability, using methods such as attention visualization, causal analysis, and rule extraction to enhance the transparency of the decision-making process, providing support for policy auditing, human-agent collaboration, and safety supervision. The fourth is the application and deployment of large-scale MARL, focusing on the problems of training efficiency, communication overhead, and scalability brought by the sharp increase in the number of agents and state dimensions, exploring mechanisms such as hierarchical structures, population modeling, and parallel training. The fifth is multi-agent safe reinforcement learning, starting from constraint satisfaction, risk control, and robustness, studying safe decision-making under adversarial perturbations, uncertainties, and policy games. Finally, this paper, combining two typical application scenarios of cooperation and competition, discusses the challenges faced by MARL in its deployment in real systems, such as insufficient sample efficiency, difficulty in simulation-to-real transfer, and insufficient analysis of fairness and steady-state games, aiming to provide a systematic reference for the subsequent theoretical research and engineering applications of MARL.

Key words: multi-agent reinforcement learning; Markov games; large language model; meta-learning; explainability; multi-agent safe reinforcement learning

Foundation Item(s): National Natural Science Foundation of China (No.U22A2011)

1 引言

近年来,强化学习(Reinforcement Learning,RL)[1,2]作为智能体基于环境交互实现最优策略学习的核心方法,已在博弈对抗[3~5]、机器人控制[6~8]、自动驾驶[9~11]等领域展现出卓越的性能. 单智能体强化学习(Single-Agent Reinforcement Learning,SARL)模型通过马尔可夫决策过程(Markov Decision Process,MDP)对任务环境建模,依托策略优化机制驱动智能体以最大化期望奖励为目标进行学习[12,13]. 尽管SARL在许多标准任务中取得了良好效果,但在面向现实复杂环境的任务中,其能力正面临严峻挑战[14,15],尤其在交通系统、物联网、城市治理等实际应用中,智能体往往并非单独存在,而是以协同、竞争、混合的方式与其他智能体共同决策,其策略制定不可避免地受到其他智能体行为、通信限制、非稳态性等因素的影响,传统的SARL方法难以满足多智能体系统的动态决策需求[16~20].

多智能体强化学习(Multi-Agent Reinforcement Learning,MARL)应运而生[21]. 作为面向群体智能体在复杂交互环境中学习的核心范式,MARL能够在状态观测受限、行为依赖增强的背景下,建模个体行为与系统行为之间的耦合关系,学习到多智能体间协同或博弈的最优策略组合[22]. 近年来,伴随深度神经网络技术的演进与强化学习理论的完善,MARL在算法体系、理论研究和应用场景方面取得了显著进展,并逐渐成为智能控制与自主决策领域的重要支撑技术之一[23~25].

然而,MARL的研究仍面临诸多关键性挑战. 一方面,受限于多智能体之间的动态策略调整,MARL所处的环境表现出显著的非平稳性,导致策略收敛难度剧增[26];另一方面,多智能体系统中普遍存在部分可观测性与信用分配问题[27,28],即单个智能体无法明确感知整体环境状态或自身行为在全局任务中的贡献. 此外,MARL系统部署面临通信受限、智能体异构、训练收敛慢等工程瓶颈,其在真实复杂场景下的可扩展性、安全性和迁移性亦亟待提升[29~34].

更为重要的是,随着人工智能进入以"大模型协同"和"人机混合智能"为特征的新阶段,MARL面临着范式重构的机遇与挑战. 一方面,大语言模型(Large Language Model,LLM)正在改变任务建模与交互表达的方式,为MARL带来语义推理、策略抽象的新能力[35,36];另一方面,元强化学习(Meta-RL)[37,38]、因果强化学习(Causal RL)[39,40]、图神经网络(Graph Neural Network,GNN)[41,42]等技术的融合,亦正在推动MARL由"局部优化"向"可解释、泛化、群体协同"的方向迈进. 这一趋势使得MARL不仅是RL的延伸,更是智能体自治决策体系的核心组成部分.

尽管已有多篇综述论文系统回顾了MARL的研究进展,但目前仍存在以下不足:一是,大多数综述侧重传统三类方法(值函数、策略梯度、模型驱动),对LLM、Meta-RL、可解释性等新兴方向缺乏系统梳理;二是,多数综述以算法分类为主线,缺乏对多智能体系统建模

理论和实际部署难点的深入剖析;三是,缺乏对现实系统中“算法—系统—应用”一体化落地的跨层分析与趋势预测. 基于此,本文从理论建模、算法技术、实际应用三重视角,全面综述多智能体强化学习的研究现状与发展趋势. 具体贡献包括以下4点.

(1)分析MARL的理论建模框架,对MARL中的决策结构、策略分类、交互模式进行系统化梳理.

(2)纵向剖析主流算法演化路径,涵盖值函数、策略梯度、模型驱动与融合范式,并总结其适用条件与优劣对比.

(3)横向拓展前沿研究主题,重点分析LLM、Meta-RL、可解释性与安全性等方向.

(4)面向应用挑战提出部署难点分析,结合协作、竞争等典型场景,探讨MARL在实际系统中的可行性.

本文将“策略更新机制”作为第2章核心方法与技术和第3章MARL的前沿突破的划分主轴,也就是在给定数据与目标下,策略如何被直接优化与生成,这构成了MARL的内核. 具体而言,值函数方法以价值近似驱动决策,策略梯度方法以可微目标直接优化参数化策略,基于模型方法则显式学习(或给定)环境或对手动态并在模型内进行规划或混合规划与学习. 三者分别对应“用价值导向决策”“用梯度直接找策略”“在可学习世界里先演练再行动”的三条互补主轴,它们决定了训练信号的来源、改进算子的形式以及计算与样本效率的基本权衡. 相应地,本文第3章“前沿突破”讨论的是在不替代上述更新算子的前提下,对内循环进行外循环增强或侧向扩展的方向,包括基于LLM的MARL、Meta-RL、MARL的可解释性、大规模多智能体系统、多智能体安全强化学习. 上述方向附着于三大内核范式之上,旨在增强学习效果、迁移效果与工程可用性,而非改变更新内核本身. 本文旨在为研究者提供一种结构清晰、内容丰富、视角多元的MARL综述性参考,助力推动该领域在理论探索与工程实践中的深度融合,文章整体结构如图1所示.

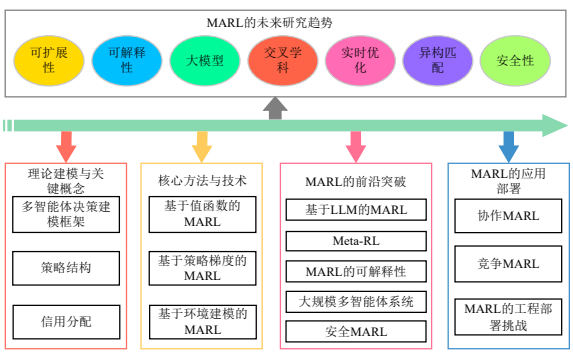

图1 文章总体架构图

2 理论建模与关键概念

2.1 多智能体决策建模框架

MARL在理论上是对SARL在多智能体环境中的扩展,其建模基础源于MDP,并进一步发展为马尔可夫博弈(Markov Game)[43]和去中心化部分可观测MDP(Decentralized Partially Observable Markov Decision Process,Dec-POMDP)[44]等形式化建模框架. 本节将系统介绍当前多智能体系统的主流建模方式,涵盖其形式定义、建模假设与适用场景.

2.1.1 Markov Game

Markov Game是对MDP在多智能体情形下的推广,是MARL的基础建模框架之一,用于描述多个智能体在共享环境中交互、竞争或协作的长期动态过程. 与传统的静态博弈不同,Markov Game考虑了状态转移的动态性,将博弈建模从单步决策推广到了多步序列决策问题[45],为MARL提供了完整的理论基础.

对于一个包含N个智能体的多智能体系统,其Markov Game可形式化定义为如下六元组:

$$\mathcal{G}=\left\langle \mathcal{S},\{\mathcal{A}_i\}_{i=1}^{N},\mathcal{P},\{r_i\}_{i=1}^{N},\gamma,\{\mathcal{O}_i\}_{i=1}^{N}\right\rangle \tag{1}$$

其中,$\mathcal{S}$是全局状态空间,描述环境的所有可能状态;$\mathcal{A}_i$是第i个智能体的动作空间,联合动作空间为$\mathcal{A}=\mathcal{A}_1\times\mathcal{A}_2\times\cdots\times\mathcal{A}_N$;$P(s'|s,a)$是状态转移函数,表示环境从状态$s$下采取动作$a$后转移至$s'$的概率;$r_i(s,a)$是第$i$个智能体的即时奖励函数,反映其在状态$s$下采取动作$a$与其他智能体联合行为后的收益;$\gamma\in(0,1)$是折扣因子,用于控制长期收益中未来奖励的影响;$\mathcal{O}_i$是智能体$i$的观测空间,反映其可见的状态信息.

Markov Game可统一描述合作、竞争、混合动机等不同关系结构的智能体群体,允许奖励函数依赖于联合动作,从而支持协同行为的隐式建模,且在理论上具备均衡性分析能力,便于从博弈论角度研究多智能体策略演化. 本文在表1中介绍了Markov Game与MDP的区别.

表1 Markov Game与MDP的区别

维度	MDP	Markov Game
智能体数量	单一	多个智能体
策略建模	独立策略	联合或耦合策略
奖励结构	独立奖励函数	个体依赖联合奖励
状态转移	依赖单一动作	依赖联合动作
稳定性	策略收敛易分析	存在非稳态问题

2.1.2 Dec-POMDP

在现实中的多智能体系统中,智能体通常无法获取环境的全局状态信息,也无法直接访问其他智能体的观测或动作,因此需要依赖各自的局部感知进行决策. 此类环境中的学习问题可由Dec-POMDP建模,是

协作型MARL常用的理论基础之一[46,47].

Dec-POMDP是POMDP(Partially Observable Markov Decision Process)向多智能体场景的自然扩展,其形式定义为

$$\mathcal{D}=\left\langle \mathcal{S},\{\mathcal{A}_i\}_{i=1}^{N},\mathcal{P},r,\{\mathcal{O}_i\}_{i=1}^{N},\{\mathcal{Z}_i\}_{i=1}^{N},\gamma\right\rangle \tag{2}$$

其中,$\mathcal{Z}_i(o_i|s,a)$是观测函数,描述环境状态及联合动作下,智能体i接收到观测o_i的概率;$\gamma\in(0,1)$是折扣因子,用于控制长期收益中未来奖励的影响.

Dec-POMDP模型具有如下核心特性:(1)分布式决策结构,每个智能体仅基于自身观测进行策略选择,适合真实部署;(2)联合动作建模,状态转移和奖励函数同时依赖多个智能体的动作,实现行为协同;(3)部分可观测性处理,天然支持传感器局限、通信限制等场景建模;(4)任务最优性统一,所有智能体共享全局奖励,目标一致,适合完全协作任务.

2.1.3　CTDE范式

在多智能体系统中,完全去中心化的学习方式虽然符合部署现实,但往往因缺乏环境信息和全局反馈而导致学习效率低、策略收敛慢、系统协同性差[48]. 为此,研究者提出了中心化训练、去中心化执行(Centralized Training with Decentralized Execution, CTDE)的学习范式,以兼顾训练效率与部署可行性,目前已成为MARL中主流的建模与训练方式之一[49~65].

CTDE的设计初衷在于解决以下三类问题:

(1)非稳态性. 在完全独立学习中,个体智能体将其他体视为环境一部分,而其他智能体策略的持续更新会破坏学习环境的马尔可夫性,导致策略训练震荡甚至不收敛[26].

(2)信用分配难题. 全局奖励下,个体难以分辨自身行为对系统表现的真实贡献[27,28].

(3)信息稀缺性. 实际执行中智能体受限于局部观测,但训练时常可访问更多信息[52].

CTDE通过在训练阶段提供集中式信息支持(如全局状态、联合动作),并在执行阶段仍保持分布式策略结构,有效解决了上述问题,提升了策略学习的稳定性与泛化能力. 在CTDE架构下,MARL被划分为训练阶段和执行阶段.

训练阶段具备以下特性:①构建全局Critic(或联合值函数)$Q(s,a)$,利用全局状态s、联合动作a、全体奖励等信息进行集中优化;②可以使用Actor-Critic框架、策略梯度、值函数分解等方法;③允许通信、共享网络参数或联合优化损失.

执行阶段具备以下特性:①每个智能体仅根据自身观测o_i执行动作决策$a_i=\pi_i(o_i)$,不依赖于其他智能体的状态、动作或策略[53];②保证系统具备良好的实际部署能力,适应无通信或弱通信约束场景[54].

2.2　策略结构

在MARL中,策略的结构设计和行为模式的组织方式直接决定了算法的可扩展性、学习稳定性以及执行部署的可行性. 不同智能体可能拥有不同的观测信息、行为能力与优化目标,因此其策略建模方式呈现出高度多样性,主要表现为策略是否独立、决策是否集中两个关键维度. 本节将从策略独立性与决策组织方式两个视角出发,对常见的策略结构与行为模式进行系统分析.

2.2.1　独立策略与联合策略

在MARL中,策略结构的基本形式决定了智能体之间信息依赖的紧密程度,进而影响系统的可扩展性、训练稳定性和协作效率. 根据策略是否建模其他智能体的行为或状态,可将多智能体策略划分为独立策略[66]与联合策略[67]两类.

独立策略假设每个智能体可以基于自身局部观测独立学习与决策,而不显式建模其他智能体的存在或行为. 形式上,第i个智能体的策略定义为$\pi_i(a_i|o_i)$. 该策略结构特别适用于信息受限、通信困难或系统规模较大的场景,代表了“感知—决策—执行”完全本地化的典型架构. 独立策略的优势在于结构简洁、参数独立、易于扩展,可以自然适应异构系统与异步执行,并支持灵活部署.

联合策略假设所有智能体共享一个联合策略网络,或策略函数依赖于其他智能体的观测与动作信息[49~65],即表述为$\pi(a|s)$. 该结构强调个体间行为的联合优化,在高度协作或集中控制任务中具有较强的表现力. 联合策略的优势在于可显式建模智能体间的依赖关系,可学习更高质量的协同行为,并适合建模非线性协同结构与复杂任务目标,有利于博弈均衡分析与全局最优策略学习.

2.2.2　集中式策略和分布式策略

集中式策略假设系统拥有一个统一的控制者或中央策略网络,能够访问全局状态信息(如所有智能体的观测、动作和奖励),用于集中计算联合策略[49~65,68]. 该结构适用于完全可观测、任务紧密协作或资源高度集中的环境,可直接优化全局目标并实现协同行为的显式建模. 集中式策略的优势在于能够精确建模所有智能体间的交互,容易施加全局约束,有利于目标一致性维护,并可用博弈分析工具对系统行为进行均衡性分析.

分布式策略强调智能体基于自身局部观测独立作出决策,无需依赖其他智能体的状态或行为. 该策略结构广泛适用于通信资源受限、异构系统部署与分布式自主控制任务,具备较强的现实可行性与扩展性[69]. 分布式策略的优势在于部署灵活,适配异构平台、边缘设

备与低通信场景,支持异步更新与局部响应,提高系统鲁棒性,并具备较强的系统扩展性.

2.3 信用分配

在MARL中,多个智能体共同影响环境状态的演化和全局奖励的生成. 当系统采用统一的全局奖励函数时,每个智能体很难直接得知自身行为对整体奖励的实际贡献,导致学习信号模糊、策略更新不明确. 为此,研究者提出了信用分配机制[27,28,59],旨在将全局奖励合理拆解并反馈至各智能体,从而提升策略学习的效率与协同性.

2.3.1 问题定义与挑战

信用分配问题指的是在多智能体系统中如何准确识别并量化每个智能体的行为对整体任务结果的边际影响. 该问题在协作型任务中尤为关键,例如在机器人团队、无人机编队等需要多个体协同完成目标的场景中. 信用分配主要面临的挑战包括以下4点:

(1)奖励共享导致学习信号模糊. 个体无法区分个体行为与全局奖励之间的因果关系.

(2)策略间耦合性强. 其他智能体行为也影响全局奖励,造成个体优化方向不清.

(3)大规模系统扩展难. 随着智能体数量增加,计算个体贡献的复杂度显著上升.

(4)训练阶段信息不完备. 若系统部分可观测,则无法准确反推出智能体的边际效应.

2.3.2 主流信用分配方法

为应对上述挑战,近年来研究者提出了多种有效的信用分配机制,主要分为以下4类.

(1)差分奖励机制

差分奖励方法[70]通过构造一个对照系统,计算在移除当前智能体行为后的奖励差异,衡量其边际影响. 第i个体的差分奖励定义为

$$D_i = r(s,a) - r(s,a_{-i}) \tag{3}$$

其中,a_{-i}表示将智能体i的动作替换为默认动作后的联合动作.

(2)值函数分解机制

该类方法通过将全局Q值函数Q_{tot}分解为个体Q值Q_i的组合结构,实现对总体Q值的局部化归因. 例如,在VDN[55]中,全局Q值与个体Q值的关系表示为

$$Q_{\text{tot}} = \sum_{i=1}^{N} Q_i(o_i, a_i) \tag{4}$$

在QMIX[56]中,全局Q值与个体Q值的关系表示为

$$Q_{\text{tot}} = f(Q_1, Q_2, \cdots, Q_N, s) \tag{5}$$

(3)反事实优势估计

反事实优势估计方法基于策略梯度方法,在计算每个体策略梯度时引入反事实基线,即考虑该体行为变化对联合动作结果的影响,例如COMA算法[59]的优势函数的计算方式为

$$A_i(s,a) = Q(s,a) - \sum_{a_i'} \pi_i\left(a_i' \mid o_i\right) Q\left(s,(a_i', a_{-i})\right) \tag{6}$$

(4)可学习的信用分配结构

近年来,研究者提出将信用分配过程本身作为一个可学习的神经网络模块,引入注意力机制、GNN或角色建模机制对个体影响力进行自动估计. 例如,在QPLEX[58]中,利用优势重分配函数对个体贡献进行非线性加权;在ROMA[62]中,通过角色编码网络引导个体学习不同的信用归因方式.

2.4 本章小结与思考

本章围绕MARL的关键建模框架进行了系统分析,重点探讨了Markov Game、Dec-POMDP与CTDE这3种主要的建模方法. 通过对比分析,每种建模框架的选择都受到不同任务需求和假设条件的约束.

Markov Game提供了较为简洁的理论框架,适用于较为传统的环境中,特别是状态和动作较为明确且可观测的场景. 然而,在处理较为复杂的协作任务时,它可能面临较为严重的策略收敛问题,尤其是当智能体数量增加时,协作行为的建模将变得更加困难.

Dec-POMDP为部分可观测的环境提供了自然的扩展,适合处理信息不完全的协作任务. 其优势在于能够有效应对现实环境中的信息限制,但其复杂度和计算需求也显著提升,且训练阶段的信息不完全性可能导致性能的波动.

CTDE是当前颇为主流的训练方式,能够在训练过程中使用全局信息提升学习效率,而在执行阶段保持智能体的独立性. 这种方法特别适合需要高效训练但又要求部署可行的任务. 然而,CTDE在实际部署时存在内在的不一致问题,尤其是在集中式Critic与去分布式Actor之间信息传递不一致时,可能导致策略漂移和执行不稳定.

尽管MARL框架在理论上已提供了坚实的基础,但随着任务的复杂性和智能体规模的扩大,现有模型在训练与执行一致性、系统扩展性以及信用分配等方面仍面临显著挑战. 特别是在实际应用中,训练阶段和执行阶段的信息不一致时常导致策略漂移,影响系统的稳定性与泛化能力. 未来的研究需要突破这些瓶颈,探索更高效的信息传递机制和训练方法,尤其是通过引入信息瓶颈或执行一致性的约束来保持训练与执行的一致性. 此外,随着智能体数量的增加,以上3种架构在大规模系统中的计算复杂度和训练难度愈加突出,如何利用分布式学习和GNN等技术来提升系统的扩展性和训练效率,将是未来的关键问题. 最后,信用分配问题依然是多智能体系统中的核心挑战之一,未

来可能通过反事实学习和元学习等方法,进一步精细化和灵活化信用分配机制,从而提升协作效率和训练稳定性. 这些方向将为MARL的发展提供更强大的理论支撑和实际应用能力.

3　核心方法与技术

3.1　基于值函数的MARL

基于值函数的方法[55,56]是MARL中最早和最常见的策略之一. 值函数用于估计每个智能体在给定状态下采取某一行动的价值,即期望的累积奖励. 这些方法基于对环境状态和动作空间的评估,以学习最优的策略来最大化智能体的期望奖励. 通过对智能体动作价值的评估,值函数方法为系统的学习和决策提供了理论依据.

在MARL中,值函数方法通常涉及智能体的状态价值函数和动作价值函数,分别表示给定状态下的期望奖励和给定状态—动作对下的期望奖励. 对于单一智能体的强化学习,Q-learning[71]和SARSA[72]等方法利用动作价值函数$Q(s,a)$来指导智能体决策. 然而,在多智能体环境中,由于智能体之间的相互影响,值函数的设计和更新变得更加复杂. 图2展示了经典的基于值函数的MARL算法的发展时间线.

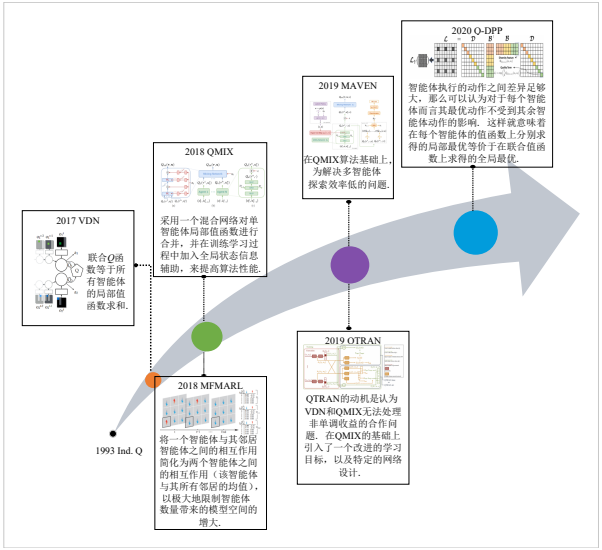

图2　基于值函数的MARL经典算法时间线

3.1.1　独立Q学习

独立Q学习(Independent Q-Learning, IQL)[66]是基于值函数的经典MARL算法之一. 在IQL中,每个智能体假设其他智能体的策略是固定的,独立地维护个体Q值函数. 每个智能体根据局部信息(即自身的状态和行为)来更新Q值函数,从而在迭代中逐步逼近最优的动作. 然而,IQL在多智能体系统中的应用存在明显的局限性. 由于每个智能体的Q值更新是基于其他智能体策略稳定的假设,这种方法忽略了环境的非平稳性,导致在实际环境中,Q值的更新往往不稳定,且不能保证收敛.

尽管IQL存在局限性,但它的简单性和可扩展性使其在许多小规模和较为简单的任务中仍然有效. 为了弥补IQL的不足,研究者们提出了更复杂的值函数分解方法,以提高多智能体系统中的学习效率和稳定性.

3.1.2　VDN

VDN[55]是一种针对多智能体环境中值函数分解的创新方法. 在VDN中,系统的全局Q值函数被分解为各个智能体的局部Q值函数之和. 具体而言,假设有N个智能体,每个智能体i都维护一个局部Q值函数$Q_i(s,a)$. 系统的全局Q值函数则被定义为各个智能体局部Q值的和.

从训练机制看,VDN以全局时序差分误差作为监督信号,但梯度通过可将分解自然地分摊到各$Q_i(s,a)$. 这种"结构化信用分配"弱化了纯粹独立Q学习中的非平稳性与信用错配问题,并常配合参数共享与角色标识以提高样本效率与泛化能力. 由于求和结构保持了单调性,目标网络与经验回放的数值稳定性相对较好,实践中易于调参、收敛速度较快.

VDN的代价是表达能力受限. 线性分解隐含各主体贡献的独立性假设,无法精确刻画依赖联合动作配置的条件性协同、抑制关系与角色切换等非线性耦合. 例如"协同门控"或"二者同时行动才有回报"的场景,本质上需要一个显式的交互项$h(a_1,a_2,\cdots,a_n)$. VDN只能以各$Q_i(s,a)$的线性叠加去近似这类协同,往往在只部分满足条件的状态上产生系统性偏差,进而诱发"偷懒智能体"等次优均衡. 在部分可观测与异构拓扑下,这一表示缺口会放大,导致对分布外结构与规模变化的外推能力下降. 这促使后续研究提出了更加复杂的值函数分解方法,如QMIX等.

3.1.3　QMIX

QMIX[56]是另一种基于值函数的MARL方法,它通过非线性函数的组合来分解全局Q值. 与VDN不同,QMIX不仅将局部Q值函数加和,还通过一个神经网络将局部Q值函数非线性地混合,从而得到全局Q值函数:

$$Q(s,a)=f\left(Q_1(s,a_1),Q_2(s,a_2),\cdots,Q_N(s,a_N)\right) \tag{7}$$

其中,f是一个单调递增的神经网络,用以学习将局部Q值拟合为全局Q值. QMIX相比于VDN的优势在于其能够通过f表达智能体之间的复杂互动,从而解决了VDN的线性加法而导致的拟合精度不稳定的问题. QMIX的关键是如下所示的单调性约束:

$$\frac{\partial Q_{\text{tot}}}{\partial Q_i}\geqslant 0,\quad i=1,2,\cdots,N \tag{8}$$

QMIX之所以需要满足单调性约束，是因为CTDE框架要求训练期可用全局信息，而执行期各智能体必须独立贪心决策. 为保证分布式贪心与集中式联合贪心一致，需要满足IGM性质：

$$\begin{aligned}&\arg\max_{a} Q_{\text{tot}}(s,a)\\&=\Big(\arg\max_{a_1} Q_1(s,a_1),\arg\max_{a_2} Q_2(s,a_2),\\&\quad\cdots,\arg\max_{a_N} Q_N(s,a_N)\Big)\end{aligned}\tag{9}$$

其中，式(8)是实现式(9)的充分条件. 其含义是，任何一个局部价值上升都不会使全局价值下降，由此可以把指数级的联合动作搜索化为线性的局部选择，确保去中心化执行可行.

单调性带来两点直接后果：①可执行性得到结构性保证，训练所得的分解在执行期天然对齐；②可表示的全局价值函数族被限制在满足IGM的那一类，更适合正协同任务与多数合作场景. 若存在强互斥或抑制关系，例如“二选一”型依赖，则最优联合动作往往不是由各体的局部最优拼接而成，此时QMIX难以精确表达，容易出现信用分配偏差与性能上限. 并且，QMIX通过超网络输出混合网络权重，并以非负化等手段保证式(8). 从梯度角度看，$\partial Q_{\text{tot}}/\partial Q_i \geqslant 0$使局部价值的贡献权重天然非负，有利于稳定的正向信用分配. 同时，也难以直接表示“负贡献”或抑制性相互作用，这在表示能力上构成内在边界.

QMIX在许多合作任务中表现出了良好的性能，尤其是在多智能体系统中. 当任务复杂且涉及智能体间的非线性关系时，QMIX通过更为精细的值函数组合，能够更有效地协调智能体行为，提高整体系统的效能. 在大规模合作任务、资源调度和分配问题中，QMIX已被证明能够显著提升任务完成度.

3.1.4 QTRAN

在QMIX基础上，QTRAN[57]旨在弥补基于值函数的MARL的单调性分解的表示边界. 其核心并非简单“变换局部Q值”，而是同时学习3个对象：集中式的全局$Q(s,a)$、每个智能体的局部效用$U_i(\tau_i,a_i)$，以及一个与状态相关的势函数$V(s)$. QTRAN在一个“变换空间”中通过两类一致性约束，把全局价值与可分解的局部效用软性对齐，从而在不要求单调性的前提下，仍然保证去中心化贪心执行的可行性.

一致性约束的直观含义如下.

(1)最优一致性. 用集中式Q选出的联合最优动作a^*应与可分解效用的联合贪心一致，对应关系为

$$Q(s,a^*)=\sum_i U_i(\tau_i,a_i^*)+V(s)\tag{10}$$

(2)非最优上界. 任意联合动作a都应满足：

$$Q(s,a)\geqslant\sum_i U_i(\tau_i,a_i)+V(s)\tag{11}$$

前者把“全局最优”映射到“局部最优的拼接”，后者把可分解效用的和约束为集中式Q的下界，从而避免由局部贪心产生的伪最优. 训练时以两项惩罚损失软性逼近上述等式与不等式，再叠加时序差分损失更新集中式Q与各U_i. 这套机制可以证明能恢复去中心化可执行性，同时把可表示的联合价值扩展到包含强非线性协同与互斥关系的情形，例如“二选一”“异步互补”等.

与QMIX的对比可概括为两点：

(a)在表达力方面，QMIX依赖单调性保证IGM性质，表达上偏向“个体改进带来团队同向改进”的协同结构；而QTRAN削弱了单调性限制，通过势函数加上界约束，允许全局最优并非各体局部最优的逐点拼接，能覆盖更一般的博弈耦合与抑制关系.

(b)在优化代价方面，QTRAN需要同时学习Q、$\{U_i\}$、V并平衡多项一致性损失，训练对采样质量、系数权重与函数近似误差更敏感，样本效率与稳定性常低于带强结构偏置的QMIX.

此外，QTRAN在分解Q值函数时，智能体不仅考虑局部Q值，还要确保所有智能体的Q值变化遵循某种关系，防止局部更新的不同步导致全局奖励的失衡. QTRAN的优势在于，它能够处理复杂的多智能体互动，尤其是在智能体之间的协作和资源共享过程中，能够保持各智能体学习的平衡和一致性. 这使得QTRAN在许多动态和高维任务中，能够有效提高学习效率，并确保系统的稳定性.

从信用分配角度看，QMIX的非负混合天然对应“正向贡献汇聚”，而QTRAN通过在一致性约束下反演出各U_i的边际效用，其“谁在何时应当受罚或受奖”的信号来源于集中式Q与可分解近似之间的张力，因此能在存在互斥资源、角色切换、竞争协作并存等情形下给出更符合全局最优的分配方案. 但由于U_i的可辨识性并非唯一，若一致性惩罚过弱或函数近似容量不足，则可能出现多解与训练震荡，需要在实现中采用权重退火、动作采样增广与目标网络稳定化等工程手段.

3.1.5 QPLEX

在QMIX基础上，QPLEX[58]的设计选择可以概括为“保留IGM等价，提升单调分解的表达力”. QMIX的核心是单调混合：$Q_{\text{tot}}=f_\theta\big(s,\{q_i\}\big)$，对每个个体效用$q_i$的偏导非负. 该约束保证集中式贪心与分散式逐个体贪心一致(IGM等价)，从而支持CTDE训练与分散执行. 但代价是表达受限，即混合权重仅随状态变化，难以刻画依赖于具体动作的协同与抑制关系，角色切换与“条件性配合”也不易表示.

QPLEX正面回应了这两个痛点，在不破坏IGM的

前提下,把“单调性”从效用层挪到优势层,并让权重对动作敏感. 其基本形式为

$$q_i(\tau_i,a_i)=v_i(\tau_i)+A_i(\tau_i,a_i),\quad \sum_{a_i}A_i(\tau_i,a_i)=0 \tag{12}$$

$$Q_{\text{tot}}(s,\tau,a)=V_{\text{tot}}(s,\tau)+\sum_i\lambda_i(s,\tau,a_i)A_i(\tau_i,a_i) \tag{13}$$

其中,$\lambda_i\geqslant 0$由超网络生成.

这样做有3层含义:

(1)优势分解把“相对贡献”从“绝对效用”中分离出来,信用分配更清晰,训练方差更低.

(2)λ_i对动作可变,使系统能根据当前联合动作的组合调节个体优势的放大或抑制,从而表达条件性协同与角色切换,同时因$\lambda_i\geqslant 0$保持单调梯度符号,IGM等价仍然成立.

(3)若令λ_i与动作无关且只随状态变化,并把优势退化为效用,便可退化回QMIX,因此QPLEX严格包含QMIX的函数簇,缓解后者的表示缺口.

从“为什么”的角度看,QMIX的单调约束保证了可执行性,但把所有交互都投影到“状态依赖的全局权重”上,导致两类系统性偏差:①当最优联合动作需要对某个体的高效用做抑制时,单调结构会高估该联合动作;②当协作关系随动作翻转时(典型如角色交替),状态静态权重难以响应. QPLEX的“优势层单调+动作条件权重”正是针对这两类偏差的定点改造,既不放弃单调性带来的IGM可执行性和稳定训练,又把表达重点转移到“谁在当前动作组合下应被放大”,因此在利用QTRAN的软一致性约束与额外对齐损失的情况下,显著扩大了可分解联合价值的覆盖范围.

QPLEX相比于QMIX的优势可总结为3点:理论上,QPLEX在IGM可表示类内更接近“完备表示”,对已知的QMIX反例具备更强的拟合能力;方法上,引入双重dueling结构(个体层与混合层)后,优势成为主要的信用载体,超网络只需学习“何时放大谁”,优化更直接;实践上,QPLEX保留了QMIX的CTDE训练范式与超网络条件化思想,因而迁移成本低,却在需要条件性协同、角色切换和非加性协作的任务上获得更稳定的提升.

总体来看,基于值函数的MARL在“可执行性—表达力—可扩展性”之间形成了清晰但仍未完全闭合的三元权衡. 其优势在于以可验证的目标分解和集中训练分布执行范式,提供了稳定的训练接口与明确的贪心执行规定;其短板则集中在表达受限、训练非平稳与规模化成本等方面.

(1)表达力与可执行性的结构性边界仍然存在. 以VDN的线性可加分解与QMIX的单调混合为代表,主流方法通过个体效用到全局价值的可分解映射,确保局部贪心与全局贪心一致,从而支持分布式执行. 这一“IGM等价”带来强可执行性,却天然排除了大量需要条件性协同、角色切换或抑制关系的联合价值结构,形成表示缺口. QTRAN通过在变换空间引入一致性等式与不等式放松单调性,扩大了函数范围,但引入了附加对齐损失,训练更敏感. QPLEX把单调性从效用层转移到优势层,并让权重对动作条件敏感,在不破坏IGM的前提下增强了对条件性协同的刻画. 然而,何种联合价值类在不牺牲分散可执行性的约束下可被完备表示,仍缺乏统一的刻画与判定准则.

(2)非平稳性与信用分配导致的训练难点尚未根除. 多智能体并发更新使目标函数随时间漂移,经验回放与时序差分带来的自举偏差在协作场景中被放大,易出现价值高估、过拟合局部协调模式与训练振荡. 值分解虽然在一定程度上缓解了信用分配压力,但分解本身并非唯一,等价的全局价值可对应多组个体效用,存在可辨识性风险与“表征漂移”. 当观测为局部且含噪时,个体效用学习会混入同伴策略的偶然相关,形成伪因果,降低策略在分布外拓扑与规模上的稳健性.

(3)规模化下的计算与通信代价形成新的瓶颈. 集中式Critic或混合网络随智能体数量、动作基数与状态维度上升而急剧增大参数与内存占用,联合动作空间的指数级增长使训练目标与目标网络更新难以保持数值稳定. 基于CTDE的MARL架构在工程落地时还需要承担全局状态汇聚与回放同步的通信成本,易与实时性需求冲突. 即便采用图结构、稀疏注意与均值场近似,如何在保持表达力的同时获得可证收敛与线性甚至亚线性复杂度的训练、推理界,仍缺乏系统化答案.

由此可得到3点研究启示:

(1)应从函数类角度明确“IGM约束下的可表示边界”,给出充要条件与最小结构,避免仅以经验反例推动算法演化.

(2)需要在值分解中引入可辨识性与稳健性原则,例如以因果一致性、优势稀疏性或不变特征为正则,降低伪相关引发的信用错配与过拟合.

(3)规模化训练应把通信与计算开销纳入目标,联合设计分解结构、消息编码与更新节奏,在理论上同时给出误差—复杂度—带宽的多目标界,从而为工程部署提供可执行的资源—性能折中.

综上,值函数方法已在CTDE框架下建立了可执行与可训练的基座,但其理论边界、训练稳健性与规模化可落实性仍待进一步收紧. 围绕表示完备性判定、信用分配的可辨识化以及通信计算一体化优化的系统化推进,将决定该分支在更复杂、开放与异构的实际场景中的持续生命力.

3.2　基于策略梯度的MARL

基于策略梯度的MARL算法是MARL中的另一类重要分支,它通过直接优化智能体的策略函数来最大

化累积奖励，而不是像值函数方法那样依赖于状态或动作的价值估计[73]. 基于策略梯度的方法通常使用梯度上升或下降的方法来更新策略参数，以优化策略在每个时间步的行为. 与基于值函数的算法相比，基于策略梯度的方法在处理大规模、高维动作空间和连续动作空间时表现出较大的优势.

在SARL中，策略梯度方法已广泛应用于各种任务，如经典的DDPG[74]. 然而，MARL环境中的策略梯度方法面临的挑战更加复杂，因为每个智能体的行为不仅依赖个体策略，还受到其他智能体策略的影响，导致环境的动态性和非平稳性.

策略梯度方法的核心思想是，通过对策略函数进行参数化，并使用梯度上升或下降法来直接优化策略函数. 对于一个给定的策略$\pi_\theta(a|s)$，其中θ表示策略的参数，a是行动，s是状态，目标是最大化当前策略的期望奖励：

$$J(\theta)=E_{\pi_\theta}[R]=E_{\pi_\theta}\left[\sum_{t=0}^{T}\gamma^t r_t\right] \tag{14}$$

其中，R是智能体的累计奖励，r_t是在时间步t获得的即时奖励，γ是折扣因子. 通过对策略的参数θ计算梯度，可以使用梯度上升法来更新策略参数，从而使期望奖励最大化. 为了计算策略的梯度，通常使用蒙特卡洛法或者时序差分法. 通过采样智能体的轨迹（状态、动作和奖励序列），可以估计梯度并更新策略参数.

为了应对MARL中的非平稳性问题，研究者们提出了一些改进的基于策略梯度的MARL算法. 例如，COMA[59]和MADDPG[60]是两种常见的策略梯度方法. 具体来讲，MADDPG是在“集中训练、分散执行”的框架下，将确定性策略梯度扩展到多智能体环境的一类方法. 在训练阶段，为每个智能体配置本地Actor，同时引入以联合观测与联合动作为输入的集中式Critic，用于评估各智能体动作在当前多方交互下的价值，从而在梯度估计中显示感知队友与对手的策略变化，缓解多智能体并行学习导致的非平稳性. 在执行阶段，仅保留各自的Actor基于局部观测独立决策，无需全局信息，满足实时与去中心化要求.

MADDPG的更新机制可表述为第i个智能体的策略参数按确定性策略梯度更新，梯度项由本地Actor的导数与集中式Critic对该智能体动作的偏导相乘给出. Critic采用时序差分回归联合动作价值，目标值由即时回报和目标网络给出的下一时刻联合动作价值构成. 经验回放与双网络软更新用于稳定训练，联合动作由各智能体的目标Actor生成. 这样做使得策略优化在连续动作空间中保持高分辨率的梯度信号，同时通过集中式评估降低策略互相干扰带来的方差.

MADDPG的主要优势体现在两点：一是，集中式Critic显式条件于多方信息，使得策略更新能够适配其他智能体策略的变化，显著缓解独立学习的非平稳难题；二是，确定性策略梯度天然适配连续动作与高精度控制，避免离散化带来的近似误差. 与此同时，该方法也暴露出若干边界与代价. 集中式Critic的输入维度随智能体数量与动作维度快速增长，带来计算与样本效率压力. 信用分配未被显式建模，协作的边际贡献主要依赖隐式学习，易出现协同不稳与探索效率下降. 重放数据与当前策略存在分布偏移，策略更新的稳定性需要较强的目标网络与小步更新来维系. 在仅有局部观测的场景中，全局状态缺失会影响Critic的辨识能力，进一步放大对样本数量的需求. 这些特征决定了MADDPG适合中等规模、连续控制且通信受限可用的协作或混合型任务，而在超大规模与强部分可观测环境中往往需要结合值分解、注意力或GNN等机制，以提升可扩展性与稳健性.

在同一套CTDE框架下，MADDPG虽然使用集中式Critic辅助训练，但其策略梯度仍以“联合价值”作为信号，无法区分个体动作对团队回报的边际贡献，信用分配不清. 直观后果是：当团队只得到全局回报时，单个智能体的梯度容易受到队友动作波动的干扰，方差大、更新噪声重，出现“错奖错罚”和学习不稳定，尤其在协作且离散动作较多的任务中更为明显. COMA[59]正是为解决这一信用分配难题而提出的方法. 该方法在训练期保留集中式Critic，估计联合动作价值$Q(s,a)$，同时为每个智能体构造“反事实优势”，把队友动作固定，仅比较该智能体当前动作与其策略下的期望动作，对应的优势定义为

$$A^i(s,u)=Q(s,u)-\sum_{a_i\in A_i}\left[\pi^i(a_i|\tau^i)Q\left(s,(a_i,u^{-i})\right)\right] \tag{15}$$

这样得到的梯度信号只反映“我这一步到底帮了团队多少”，直接缓解了信用分配不明导致的高方差与错导向问题. Actor仍基于局部信息执行，满足分散执行的需求. 与MADDPG相比，COMA的改进体现在3点上：①优势估计围绕个体边际贡献构造，信用分配清晰，梯度方差显著降低；②集中式Critic在训练期统一感知多方交互，减弱由对手与队友策略变化带来的非平稳性；③保持执行期的去中心化形态，便于工程落地. 其适用边界也较明确：Critic需要处理联合动作，维度随智能体与动作规模上升而增大；反事实优势对离散动作需做一次边缘化，动作空间很大或连续时需要近似；若全局状态不可得或Critic估计偏差较大，优势也会失真. 总体而言，可以将二者的关系概括为MADDPG提供了“集中评估加分布式执行”的基本“骨架”，但在团队协作场景下欠缺精确的个体信用刻画；COMA在相同“骨架”内补上了这一块关键“筋膜”，以反事实优势把联合价值拆解为可学习、可归因的个体学习信号.

在MADDPG和COMA之后,实践表明"集中式评价加反事实基准"虽然缓解了信用分配,但训练仍易受非平稳与高方差影响,且主要面向离散动作,对评价器精度依赖较强. 为减轻这些问题,MAPPO[61]沿用CTDE思路,但以$Q(s,a^1,a^2,\cdots,a^N)$提供更稳健的优势估计,并引入近端策略优化的剪切目标抑制过激更新,从机制上提高训练可复现性与收敛稳定性.

MAPPO的核心做法是用集中式Critic计算优势A_t(常配合广义优势估计),Actor端按如下剪切目标更新:

$$L_{\text{clip}}=E\left[\min\left(r_tA_t,\text{clip}(r_t,1-\varepsilon,1+\varepsilon)A_t\right)\right] \quad (16)$$

其中,r_t为新旧策略概率比,ε为步长阈值. MAPPO通过参数共享与身份编码在同构体间复用表达,同时允许连续或离散动作,适配部分可观测场景的循环策略. 与COMA相比,MAPPO不再显式构造反事实基准来精细分摊团队回报,而是依赖更稳定的优势估计与"近端"更新来降低策略漂移与方差,因而在中大规模协作基准上表现更平滑、更易复现.

但MAPPO的代价在于其仍属on-policy,样本效率不及强off-policy方法;集中式Critic对全局或近全局信息有依赖,真实系统需配套状态重建与通信压缩;未直接解决细粒度信用分配,复杂协作下可结合注意力Critic、反事实基准或值分解进一步强化. 整体上,MAPPO可被视为在COMA之后对"稳定训练与可扩展实现"的工程化加强,它保留CTDE的信息优势,用剪切更新替代反事实校正,牺牲部分样本效率换取显著的稳定性与跨任务可复现性,使前文的MADDPG、COMA系列在更复杂环境中具备可用的"强基座".

为进一步提升策略学习在异构智能体系统中的稳定性,文献[64]提出了HAPPO算法. 该方法通过序列化更新机制避免了多个智能体并行更新所带来的干扰问题,同时保留了PPO在稳定性和样本效率方面的优势. HAPPO特别适用于异构策略结构、多任务或具有不同能力的智能体系统,在复杂任务中展现出更强的训练稳定性和策略协调性.

总结来看,基于策略梯度的MARL以直接优化策略为主线,通过参数化策略在环境交互中累计回报的梯度信息进行更新,天然适配连续动作与高维控制,并在"集中训练、分布执行"范式下用集中式评价器缓解非平稳性. 其共性机制包括3方面:①以集中式Critic聚合全局或近全局信息,产生方差更低、时序一致的优势估计,从而稳定各体策略的更新;②通过结构化的信用分配将团队回报还原到个体决策层面,提高协作学习的有效性;③在更新规则上引入稳定化手段,如近端剪切目标与广义优势估计,控制每步策略漂移,提升收敛可复现性.

代表性方法沿此主线分工互补. MADDPG保留确定性策略梯度的样本效率优势,以集中式Critic学习耦合动态,适合连续控制与局部可观测场景,但在团队回报分摊上较为粗糙. COMA在集中式评价的基础上引入反事实基准,显式度量"若仅改变某体动作,团队回报将如何变化",从而缓解信用分配与多体耦合带来的梯度偏差,提升合作任务的学习效率. MAPPO以近端策略优化为核心,结合集中式Critic和优势估计,在不依赖精细反事实建模的前提下获得更稳定的训练轨迹与更强的跨任务可复现性,成为大多数协作基准的强基座. HAPPO进一步面向异构体与并行更新,采用顺序化或分块化更新以减少相互干扰,兼顾稳定性与可扩展性.

这一系列方法围绕4个痛点给出系统化解法:用集中式评价抑制非平稳,用反事实或注意力型评价器做信用分配,用近端剪切与优势估计稳住训练,用参数共享与身份编码支撑多体与异构. 其主要收益是可在复杂耦合和连续动作中维持稳定收敛与较强的任务表现;主要代价是对全局信息与通信条件存在依赖,且多属近端更新,样本效率不及强离线策略方法;在极端稀疏奖励、强对抗博弈或超大规模系统中仍可能出现方差上升与收敛变慢. 面向工程应用,常见的有效增强路径包括以注意力或图结构的集中式Critic提升可扩展性,在策略层结合值分解或反事实基准细化信用分配,在算法层引入离线策略回放与重要性修正缓解样本效率不足,在系统层配套状态重建与通信压缩以降低对全局观测的硬依赖. 通过上述组合式设计,策略梯度路线可作为协作、博弈与快速适应任务的稳健底座,并与值分解、基于模型与安全约束等技术形成互补.

3.3　基于环境建模的MARL

MB-MARL(Model-Based Multi-Agent Reinforcement Learning)是一种基于环境建模的MARL方法[75~77]. 与传统的基于值函数和策略梯度的MARL方法不同,MB-MARL通过学习环境的转移模型和奖励模型,在模拟环境中进行规划,从而提高学习效率,尤其在稀疏奖励和复杂动态环境中具有明显优势. 它通过将环境建模与策略学习相结合,弥补了基于值函数和策略梯度方法中样本效率较低的不足. 基于模型的方法通常包括环境建模、模拟规划和策略优化3个主要步骤.

(1)环境建模

在MB-MARL中,智能体通过与环境的交互学习状态转移模型$P(s'|s,a)$. 该模型用于预测在某一给定状态s下执行动作a后环境状态s'的变化. 然后,通过学习环境奖励生成规则,建立奖励模型$r(s,a)$. 智能体能够预估执行某个动作后所获得的奖励,从而引导行为优化.

(2)模拟规划

在学习到环境模型之后,智能体可以通过在环境

模型中进行模拟，预测不同动作所带来的潜在奖励和状态转移，并根据模拟结果进行策略调整. 通过这种方式，智能体可以在有限的真实交互次数下，提前进行决策优化. 同时，智能体可以利用模型进行多步前瞻性规划，选择能够长期带来高奖励的行动路径. 此过程通过多个时间步的模拟来进行策略优化.

(3)策略优化

通过模拟规划获得的预期奖励反馈，可以指导策略网络的更新. 智能体采用基于模型的优化方法，如Q-learning、PPO或其他强化学习方法，通过环境模拟计算的奖励来优化策略.

MAMBPO[78]是经典的MB-MARL的工作，将单智能体MBPO[79]拓展至多智能体. 训练期由集中式学习器汇聚各体经验，学习一个集中式世界模型$\hat{p}_\theta$，用于预测下一时刻联合观测与奖励$o_{t+1}, r_{t+1} \sim \hat{p}_\theta(o_t, a_t)$，执行期各体基于自身观测独立行动，无需模型或中心信息. MAMBPO的优势在于执行期不依赖模型、样本效率显著提升. 其局限在于3点：①仍需集中式训练与模型学习，当环境含有外生体(如固定启发式的“猎物”)且关键变量不可观测时，模型学习易受限；②尽管效率大幅提升，现实中仍需数千回合方能达成高性能，距离“少次试验即可上线”的目标尚有差距；③MAMBPO的可扩展性在大规模系统中仍取决于集中式Critic与通信/存储开销. 工程上，建议坚持“短回滚+真实样本”混合的模型使用策略，采用模型集成控制偏差，并在多任务/多域中结合正则化与域随机化提升世界模型的稳健性.

虽然MAMBPO以“环境建模—想象滚动—策略优化”的内循环提升样本效率，但在带宽受限、邻域交互频繁的协作场景，仅靠集中训练与离线想象并不足以消解部署期的信息瓶颈与非平稳性. MAMBA[80]面向这一痛点，要求执行期允许有限邻域通信，通过在训练期构建可分发的世界模型，使每个智能体在想象空间内完成决策学习，同时将执行期的通信预算显式纳入设计，从源头兼顾样本效率与可部署性. 具体来说，MAMBA学习多智能体世界模型，使用循环隐变量结构聚合时序信息，并以注意力建模体间的相互作用，得到每个智能体的离散潜在表征与对应的观测、回报、折扣生成头. 策略与价值均在潜在空间学习，训练阶段以想象轨迹驱动近端策略优化，避免高方差的真实交互. 为稳住潜在表征，加入潜在状态与先前动作的信息约束，鼓励因果一致的表征更新；为适配数量变化，引入拓扑稀疏化约束，使模型在规模波动下保持可用.

总体而言，基于环境建模的MARL为“样本效率—前瞻规划—可部署性”提供了一条可实操的联合路径. 用结构先验与不确定性管理把模型偏差“关在笼子里”，用可扩展的潜在表示与通信设计支撑多体协作，把CTDE的训练优势延续到分散执行阶段.

3.4 本章小结与思考

本章深入探讨了基于值函数、策略梯度和环境建模这3类核心MARL方法，每种方法在不同的任务和环境下展现出了各自的优势与挑战.

基于值函数的方法，如IQL、VDN、QMIX和QTRAN，依赖于分解全局Q值来实现多智能体之间的协同决策. 尽管这些方法在结构化信用分配和可扩展性方面表现出色，但它们的表达能力依然受到限制，尤其是在面对非线性协作和复杂的动态环境时. 随着智能体数量的增加，训练的不稳定性和计算的高复杂度成为亟待解决的问题，尤其是值函数的线性或单调分解假设难以准确刻画复杂的智能体互动，未来的研究可能需要探索更灵活的非线性分解方法和更强大的模型表达能力.

基于策略梯度的方法，如MADDPG、COMA、MAPPO和HAPPO，通过集中式Critic和分布式Actor的框架缓解了多智能体环境中的非平稳性问题. 这些方法在处理连续动作空间和高维控制任务时表现出了优异的性能，但也暴露出高方差、样本效率低等缺点. 尽管有些方法引入了反事实优势或近端优化来提高稳定性，但如何进一步降低策略更新中的方差并提高训练效率，仍是未来的研究热点.

基于环境建模的方法，如MAMBPO和MAMBA，借助环境模型进行模拟规划，显著提升了样本效率，尤其在稀疏奖励和复杂动态环境下，能够有效提高学习速度. 尽管如此，这些方法依赖于集中式模型训练和模型学习，当环境动态性增加时，模型的准确性和可扩展性依然面临挑战. 未来的研究可以在优化模型学习和分布式执行能力方面做更多探索，以应对大规模系统中的挑战.

综上所述，基于值函数、策略梯度和环境建模的方法各具特色，但都面临一定的理论边界和实践挑战. 未来的研究需要关注如何平衡训练稳定性、表达能力和可扩展性，尤其是如何在多智能体系统中高效地分配信用、应对环境动态变化，并实现系统的实际可部署性.

4 MARL的前沿突破

随着MARL研究的不断深入，传统的MARL方法(如值函数方法、策略梯度方法等)在面对越来越复杂的任务时，往往面临可扩展性、泛化性和计算效率等方面的瓶颈. 为了克服这些局限，近年来，研究者们提出了几种新兴的研究方向，旨在打破传统MARL框架的限制，并拓展其应用场景. 本章将重点讨论当前MARL领域的前沿突破，具体包括LLM与MARL结合、Meta-RL的跨任务适应性、可解释性、多智能体安全强化学习等技术创新，分析其对MARL未来发展的深远影响. 本章的架构如图3所示.

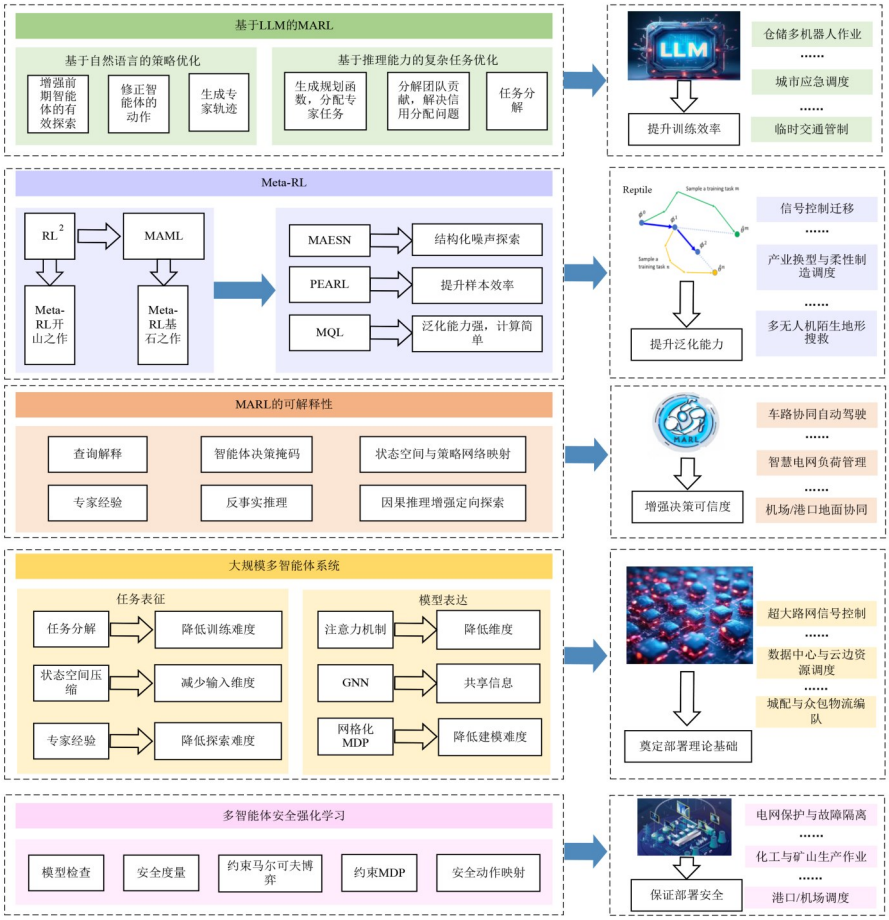

图3　本章的总体架构图

4.1　基于LLM的MARL

随着LLM在自然语言处理领域的迅猛发展,其强大的文本生成与理解能力已经引起了MARL领域的广泛关注[81~83]. 将LLM与MARL相结合,旨在通过自然语言的引导与策略生成来提升智能体的学习效率和协作能力. LLM作为一种强大的生成工具,不仅可以为智能体提供任务指导、策略协同,还能为复杂任务中的策略优化提供高效的引导. 该方法通过引入语言模型的生成能力,让智能体能够在更高层次上理解和执行任务指令,带来更灵活的策略调整和任务规划. 本节将结合当前的研究进展,详细分析LLM与MARL结合的几种重要应用方向,包括基于自然语言的策略优化和基于推理能力的复杂任务优化,并探讨每个方面的挑战与潜在解决方案.

4.1.1　基于自然语言的策略优化

LLM的最直接应用之一是在MARL的训练过程中通过自然语言输入生成策略. 通过自然语言的提示,智能体可以从LLM获取任务规划和策略建议. 这一方法的优点在于,智能体不再仅依赖于数值化的Q值或策略网络,而是通过生成的语言指令来执行任务. 这种方式不仅使策略的生成变得更加直观,还为复杂任务的执行提供了更高层次的引导.

文献[84]提出的LAMARL面向多机器人协作中的策略生成难题,将LLM与MARL紧密结合. LLM先通过指令与约束分析自动产出先验策略与奖励函数的可执行代码,并将其注入MARL训练环节. 前者为智能体提供基本可行的协作行为,后者引导高效探索与性能优化,从而减少人工设计成本、加速收敛并提升跨任务可迁移性. 与之相似,文献[85]在血管内手术的多仪器协作导航场景中提出MAFRL,以LLM提供程序化先验与上下文感知的策略指导,同时采用模糊强化学习处理LLM输出的不确定性,并将临床约束嵌入奖励函数以确保安全与合规. 二者结合既发挥了LLM的推理与知识整合能力,又保留了MARL的协作优势,最终实现更精准且可落地的导航控制.

为解决MARL中的冷启动难题,文献[86]提出了一种基于LLM的引导方法. 环境状态首先被转化为自然语言输入,LLM据此生成策略建议并输出行为动作,从而加速智能体策略收敛并减少无效探索. 同时,LLM可辅助推断其他智能体的意图,提升协作效率. 作者还

开发了一个兼容SMAC环境的工具包,支持自动提示生成、策略转换与可视化,便于集成与实验.

在智能体沟通能力方面,文献[87]提出的LangGround系统以LLM为核心,生成结构化语言轨迹并与任务环境进行对齐,进而引导MARL智能体学习符合人类习惯的通信协议.LLM通过自我博弈生成富含语义的专家轨迹,作为MARL策略学习的基础数据,有效增强了团队任务性能与可解释性.

文献[88]聚焦于多机器人系统中的自然语言导航任务.LLM将语言指令映射至潜在观测空间,并结合多机器人动作生成训练数据.通过Expected SARSA进行离线学习,该方法不仅提升了策略泛化能力,还成功部署于真实机器人,展示了语言驱动导航控制的可行性.

与此同时,LLM在多智能体通信中的作用也逐渐凸显.文献[89]提出GMAC框架,利用生成式LLM压缩观测数据.通过语义提取与预测,LLM将冗余状态信息转化为紧凑的语言描述,并在接收端重建关键信息,显著减少通信负担并加快策略收敛速度.

综上所述,LLM在MARL中的应用不断拓展,不仅助力任务分解、协作引导与语言对齐,还在通信优化与策略冷启动中展现出巨大潜力.未来工作可进一步探索LLM在因果建模、多模态交互及实时反馈机制中的作用,推动构建通用、可解释、高效的多智能体系统.

4.1.2 基于推理能力的复杂任务优化

LLM能够以任务降解、状态理解与奖励设计等方式辅助MARL的策略学习,从而提升训练效率与决策精度,特别是在复杂环境、稀疏奖励与部分可观测场景中表现出显著优势.文献[90]提出的TALKER面向无人机集群的长期任务规划与执行,采用分层交互框架由LLM负责高层的任务分解与目标分配,MARL承担低层个体技能学习(如导航、避障),并引入知识扩展与任务激活机制,使LLM能随用户反馈持续更新知识并生成更贴合任务需求的规划方案.

在更复杂任务协作与安全性挑战方面,文献[91]提出的LGC-MARL框架将LLM用于任务分解和行动图构建,生成清晰的协作目标与子任务序列.LLM还能动态设计奖励函数,鼓励多智能体系统形成稳定有效的协作行为.MARL模块则利用这些奖励与指令进行策略学习与优化.该方法实现了LLM与MARL的闭环交互,提升了复杂任务执行下的智能体学习能力与系统稳定性.

针对MARL中信用分配与部分可观察性问题,文献[92]提出了LERO框架,利用LLM生成混合奖励函数并引导信用分配过程.LLM根据任务目标与环境状态,动态推理个体贡献并生成个体化奖励信号,有效解决了传统方法中团队贡献难以量化的问题.同时,LLM可基于局部观测推断环境上下文,夯实智能体的决策基础.LERO还将LLM嵌入进化过程,不断优化奖励结构和观测增强机制,构建出稳定高效的协作学习闭环.

在城市智能调度场景中,文献[93]提出了LiMeda框架,将LLM与MARL结合,解决异构车辆资源调度与协作导航问题.LLM模块对自然语言任务描述与车辆信息进行结构化解析,并完成高效任务分配;MARL模块则执行导航策略学习,实现异构车辆的高效协作.该框架弥补了传统方法在复杂城市环境与异构资源调度中的不足,推动了多智能体系统在智能交通中的应用落地.

综上所述,LLM为MARL提供了强大的语义推理、上下文建模与任务降解能力,有效提升了智能体在动态环境中的策略质量与协作效率.未来研究可进一步拓展LLM在动态规划、语言引导式控制、因果推理等方面的能力,为构建通用、解释性强且高效的MARL系统提供更强支撑.

总体而言,基于自然语言的策略优化通过将LLM作为策略生成器,利用自然语言的引导提高了多智能体系统的协作效率与任务执行的灵活性.LLM能够根据自然语言指令动态调整智能体的行为,使得策略的生成具有高可解释性.与传统的数值化策略方法相比,LLM可以在复杂的任务和动态环境中提供更加灵活的策略调整和任务引导.基于推理能力的复杂任务优化则侧重于利用LLM的推理能力在复杂的多智能体任务中进行有效的任务规划与策略调整.通过深层次的推理,LLM能够优化任务执行过程中的每一个细节,特别是在长时序决策或环境动态变化较大的任务中.LLM的推理能力有助于智能体在复杂环境中快速适应并做出最优决策,显著提高任务的协作效率和执行效果.尽管这两个方向展示了LLM与MARL结合后的巨大潜力,但在实践中仍面临着许多挑战.如何提升LLM生成策略的准确性和一致性,如何加强推理过程的深度和效率,以及如何确保这些方法在复杂动态环境中可靠运行,都是当前研究亟待解决的问题,总结如下.

(1)生成内容的准确性与一致性

LLM在策略生成过程中,特别是在面对复杂的多智能体任务时,可能会出现生成内容的模糊性或歧义性.尽管LLM在理解和生成自然语言上有很强的能力,但其生成的内容仍然可能与实际任务的需求不完全一致,导致智能体执行的策略与预期不符.这种生成过程的不稳定性可能影响系统的表现.

(2)智能体间的协作与语言指令的协调

在多智能体系统中,各个智能体之间需要协同合作以完成任务.LLM生成的语言指令是否能够在智能体之间保持一致,成为影响系统性能的关键因素.如果

不同智能体接收到的指令不一致,可能会导致协作失败或效率低下. 因此,如何确保多智能体间的策略协同和指令的一致性,是一个需要解决的问题.

(3)环境适应性与任务迁移问题

LLM 的策略优化能力通常依赖于任务的先验知识. 然而,在面对新环境或任务时,LLM 可能无法快速适应并生成合适的策略. 尽管 LLM 在训练过程中能够处理某些任务,但其泛化能力在面对未知任务时可能受到限制,如何提升 LLM 的自适应性和迁移能力,是当前研究中的挑战之一.

(4)推理深度与多步骤决策

在处理复杂任务时,LLM 需要具备较强的推理能力,能够考虑任务的长期目标,并进行多步骤的决策. 然而,目前的 LLM 通常更擅长生成短期决策,对于长时序的任务推理深度和决策层次存在不足. 如何增强 LLM 在复杂任务中的推理深度和层次性,使其能够有效应对长期任务和复杂决策,是一个亟待解决的问题.

(5)推理效率与计算开销

LLM 的推理过程通常需要大量的计算资源和时间,而在面对复杂任务时,LLM 的推理效率可能成为瓶颈. 在多智能体系统中,随着智能体数量的增加,推理的计算需求急剧上升,可能导致系统的延迟和性能瓶颈. 如何优化推理过程,提升效率并降低计算开销,是未来研究的一个关键方向.

4.2　Meta-RL

4.2.1　Meta-RL 的概念与理论

Meta-RL 是一种基于元学习的方法,其核心目标是让智能体不仅能学习如何在一个任务中优化策略,还能学习如何快速适应新任务. 传统的 RL 方法通常假设任务和环境在整个训练过程中都是固定的,智能体需要通过与环境的多次交互来获得最优策略. 然而,Meta-RL 关注的是智能体如何从一组任务中学习,进而能够在见到新任务时迅速适应,通过最少的经验就能解决新的问题. Meta-RL 结合了元学习的思想,元学习的基本目的是学习如何学习. 在 Meta-RL 的框架下,智能体通过在多种任务环境中进行训练,使得其不仅学习到具体任务的最优策略,还能通过学习如何快速适应新任务,减少每个新任务所需的训练时间和样本量.

4.2.2　Meta-RL 的核心原理

元强化学习的基本原理可以通过以下几点来概括:

(1)元训练. Meta-RL 中的元训练阶段是通过多个任务来训练智能体的学习策略. 具体来说,智能体通过与多种环境的交互,学习到一种可以快速适应新任务的通用策略或学习过程.

(2)元测试. 在元测试阶段,智能体会面临一个全新的任务,它需要在该任务中迅速调整策略并优化表现. 元测试是验证智能体在新任务上的快速适应能力,即验证元学习的效果.

(3)快速适应能力. Meta-RL 的关键在于让智能体通过少量的数据和经验,能够在新任务中快速调整其行为策略. 这是通过在多个任务中积累的经验,使得智能体在面对新任务时具备高效的学习能力.

(4)任务知识共享. 通过元学习,智能体能够在多个任务之间迁移经验,实现知识共享,从而在新任务中更快地找到有效的策略.

4.2.3　Meta-RL 的算法与工作

RL^2作为 Meta-RL 的开山之作[94],旨在解决传统 RL 方法样本效率低的问题. 其核心思想是将 RL 算法本身视为一个学习目标,并使用标准的 RL 算法对其进行优化. 具体来说,RL^2将 RL 算法编码到一个循环神经网络(Recurrent Neural Network,RNN)中,并通过 RL 算法来学习 RNN 的权重. RNN 接收典型的 RL 算法所接收的所有信息,包括观察值、动作、奖励和终止标志,并在给定的 MDP 中保留其状态. 实验结果表明,RL^2在小型问题上取得了与理论上最优算法相当的性能. 然而,RL^2的性能受限于外层 RL 算法,可以使用更好的算法或架构来进一步提高性能. 在 Meta-RL 的众多算法中,MAML 是最具代表性的算法之一[95]. MAML 的目标是通过优化一个通用的初始模型,使得智能体能够在面对新任务时,仅通过少量的梯度更新,快速适应并获得较好的性能. MAML 的核心思想是在多个任务中进行训练,优化智能体的初始化参数,使得其能够在新任务上通过少量的学习步骤(如梯度更新)迅速适应. 具体来说,MAML 算法通过以下三个步骤进行训练和优化.

(1)任务采样. 从任务分布中采样一个任务集合,每个任务 T 都有其独特的目标和奖励结构.

(2)梯度更新. 对于每个任务 T,使用当前的模型参数 θ 进行几次梯度更新,得到新的模型参数 θ',即

$$\theta' = \theta - \alpha \nabla_\theta L_T(f_\theta) \tag{17}$$

其中,θ 是智能体的初始化参数,θ' 是经过更新后的参数,α 是学习率,L_T 是任务 T 的损失函数,f_θ 是当前任务下的模型表现.

(3)元更新. 在元训练过程中,算法通过在多个任务上计算损失,并对所有任务的更新过程进行平均,得到一个共享的更新步骤. 目标是通过优化模型参数 θ,使得模型能够快速适应新的任务.

在 MAML 框架的基础上,研究者提出了多种扩展方法以提高探索效率与迁移能力. 例如,文献[96]提出的 MAESN 算法结合了结构化噪声与 MAML,通过引入时序相关噪声建模探索空间,使智能体在面对稀疏或延迟奖励环境时,能更有效地执行有针对性的探索指令.

该算法同时优化策略参数和探索潜变量,为新任务提供结构化的探索行为,显著优于传统随机噪声方法.

进一步地,文献[97]提出的PEARL方法将隐变量建模与离线策略学习结合,通过结构化的变分推理机制刻画任务上下文分布,在少量交互数据基础上实现任务后验分布估计与策略调整.与依赖在线优化的传统方法相比,PEARL具备更高的样本效率与计算效率,尤其在高维状态空间中表现出良好的适应性与泛化能力.

文献[98]在PEARL的基础上提出了Meta-Q-Learning(MQL),该方法采用上下文变量表示任务历史轨迹信息,并设计多任务目标函数,结合离线数据执行离策略更新.通过倾向性估计机制扩展可用数据范围,MQL在保持结构简洁与计算负担较低的同时,实现了优异的样本效率与任务迁移能力.

随着Meta-RL与新兴技术的融合,研究者亦探索了跨范式的方法创新.文献[99]提出QM2ARL框架,将量子神经网络与Meta-RL结合,通过角度参数与极性参数实现分阶段策略优化.角度训练用于学习元策略,极性微调支持快速适应新任务.此外,QM2ARL引入角度—极性正则化与极性记忆机制,增强模型的泛化能力与动态适应性,在复杂多任务系统中展现出优越性能.

在理论层面,文献[100]建立了一个面向多任务MARL的集体元学习框架,通过对3类马尔可夫博弈中的纳什均衡策略进行元学习,系统分析了任务相似性对策略收敛速度的影响.研究表明,任务间相似性不仅能加速策略收敛,还能提升泛化能力.该工作设计了具备初始化收敛性保证的Meta-MARL算法,理论分析与实验验证均显示其在零和博弈与协同任务中的显著优势.

综上所述,Meta-RL为MARL提供了强大的任务迁移与适应能力支持.从MAML的模型无关性出发,研究者发展了MAESN的结构化探索机制、PEARL的离线推理优化、MQL的上下文建模策略以及QM2ARL的量子元学习方法,均有效提升了多任务环境下的泛化效率与学习速度.与此同时,理论框架的完善亦为Meta-RL在多智能体系统中的大规模部署奠定了基础.未来工作可进一步拓展至更复杂的任务切换、博弈协作与长期依赖问题上,为构建通用、稳健的智能体学习机制提供更强支撑.尽管Meta-RL在多智能体系统中的应用展现出巨大的潜力,但仍然面临若干挑战.首先,任务相似性度量仍然是一个难点,尤其是在高维状态空间和复杂环境中,如何准确计算任务之间的相似性以实现有效的迁移学习,仍需要进一步的研究.其次,样本效率仍然是Meta-RL中的一个关键问题.虽然PEARL和MQL等方法提高了样本效率,但在高维任务和多智能体系统中,如何进一步优化样本利用并减少计算开销,依然是一个挑战.再次,在动态变化的环境中,智能体如何在不同任务之间稳定迁移,并在任务间保持高效的协作,仍需要解决任务间的协作稳定性问题.最后,多智能体系统的协调与合作问题在元强化学习的框架下也显得尤为复杂,如何在多个智能体间实现快速的信息共享和协作,提升系统的整体性能,仍是未来研究的重要方向.

4.3 MARL的可解释性

随着MARL模型结构的不断复杂化,可解释性与信任性问题逐渐成为其在实际部署中的关键挑战.尽管MARL在群体决策与任务协作等方面表现卓越,但其“黑盒”特性使得智能体行为与决策过程缺乏透明度,难以理解.尤其在自动驾驶、医疗、金融等高风险场景中,开发者和用户需明确掌握智能体做出特定行为的原因,以判断其是否遵循了合理的学习策略.因此,提升MARL模型的可解释性,特别是基于决策树与符号逻辑的MARL的内在策略表示方法和基于反事实推理与因果建模的MARL策略事后解释方法,已成为推动该技术应用的重要方向.

4.3.1 MARL的内在策略解释

在MARL中,内在策略解释是在策略本身的表示里加入结构化先验,让策略从一开始就能看得懂、查得清、验得过,同时不牺牲在线响应和部署可控性.现在常用的做法是用规则、逻辑或决策树来当策略的载体,可以在CTDE框架里配合使用.训练时用神经网络学全局表征与信用分配,并加上逻辑或树等约束稳住收敛.执行时输出可读的树或规则,并用在线规则检查与动作替换守住安全边界.这样就能同时做到“性能好、说得清、可落地”.

以先天可解释方向为例,文献[101]将逻辑神经网络与概率逻辑神经网络作为核心载体,把领域规则嵌入可学习的逻辑子式,并通过概率一致地激活实现对未观测变量的推断.由此一方面以规则收缩搜索空间,改善样本效率与训练稳定性;另一方面借助概率推理,缓解部分可观测与环境不确定性带来的策略脆弱.不过,当感知维度升高、动力学更复杂时,规则的规模与维护成本迅速攀升,因而需要与特征抽象、规则稀疏化等机制配套,才能维持工程上的可持续性.

在符号策略融合方向,文献[102]通过混合状态表征承接像素与对象要素,在融合模块中合成神经网络策略与符号策略的动作分布.符号概念与评价函数由LLM生成,显著降低专家先验构建的负担,在环境变化下也更具稳健性.该体系为多智能体拓展提供了统一模板,即神经网络策略专注感知与细粒度控制,符号策略承载协作协议与角色规范,二者在融合层达成一致动作选择,从而在协作博弈中既保留表达力,又保留决

策逻辑的透明度.

另一条成熟的工程化表达来自决策树. 文献[103]将多智能体的决策过程映射为结构清晰的决策树,通过时间相关状态的聚合、同态智能体间的关系挖掘与复杂度受控的单元搜索,显式刻画层级意图与协作结构. 决策树天然适合加载规程与安全约束,延迟可控,便于审计与回归测试,但其对高维连续控制的表达力度相对不足,更适合与底层连续控制器分层对接,从而在上层保持可解释框架,在下层确保控制精度与响应速度.

面向值分解与信用分配的场景,文献[104]提出以循环软决策树学习个体的时序决策路径,并用可加混合实现联合动作值的分解. 其中,决策路径的可视化直接给出"何时、因何、如何"的因果脉络,而混合权重对应信用分配的显式刻画,再配合参数共享,还能缓解"懒惰智能体"问题. 该架构在解释性与性能之间取得良好平衡,体现了在CTDE范式下"可解释值分解"的一条务实路线.

综上所述,内在策略解释在CTDE与实际部署场景中,将神经表征、逻辑约束分层协同,并辅以在线规则检查与动作替换,可将全局学习的能力与局部决策的可审计性有机统一,为高风险、多主体、强监管的应用提供可解释、可验证且可落地的策略解法.

4.3.2　MARL的策略事后解释

MARL的事后解释聚焦于反事实推理、因果建模等方法. 文献[105]系统探讨了集中式与分布式MARL算法的可解释方法. 在集中式设定下,作者提出了策略总结、行为可视化和时间序列分析等方法以揭示策略背后的逻辑. 而在分布式场景中,针对局部观测限制,文献设计了基于Hasse图与位置叠加的策略可视化手段,有效辅助用户理解各智能体的角色分工与行为机制. 此外,文中引入的查询式解释机制,支持用户追踪模型的中间推理过程,尤其适用于决策路径复杂的任务.

进一步地,文献[106]提出了一种具有可扩展性的可解释MARL方法IMARL-MRP,聚焦于状态空间与策略网络首层权重之间的映射关系,建立了任务到策略的数学解释模型. 该方法核心在于引入"最小奖励参与机制",识别能够代表任务反馈的最少智能体,从而在网络规模变化时避免重复训练,显著提升自主水下航行器等大规模多智能体系统的适应性与迁移能力. 文献[107]则通过反事实推理引入智能体重要性评估方法EMAI,通过掩蔽目标智能体并分析其行为对奖励的影响,量化其对整体任务的贡献. 该方法基于CTDE框架,有效缓解联合动作空间维度爆炸问题,支持智能体间的依赖建模,并在任务分配与攻防策略设计中展现出良好的可解释能力.

因果推理亦成为提升MARL可解释性的重要方向. 文献[108]提出因果引导探索方法CGE,通过条件平均处理效应构建状态依赖因果度量,引导智能体优先探索对环境具有更强影响的状态,提升了探索效率,强化了智能体对反馈机制的理解,在协同任务中展现出因果推理的实际价值. 文献[109]则进一步提出基于因果建模的行动影响模型,能够生成"为什么/为什么不采取该行为"的反事实解释. 实验表明,该方法在任务预测、用户理解与解释满意度方面均优于传统奖励驱动型方法,为可解释性在现实部署中的应用提供了有效支持.

综上,当前MARL可解释性研究正在从静态行为可视化、关键智能体识别逐步迈向因果建模驱动的机制解释,其在提高用户信任度、增强模型鲁棒性与推动实际部署等方面具有重要意义. 未来研究可进一步结合图结构建模、语言生成以及专家知识引导等方式,探索多模态、可交互的可解释性MARL框架.

这些工作展示了如何通过不同的方式提高MARL的可解释性,特别是在多任务、多智能体协作的环境中,因果推理和智能体行为的可视化方法为理解智能体决策提供了有力支持. 这些创新不仅提升了模型的透明度,还增强了系统的安全性、可靠性和用户对智能体决策的信任. 尽管在MARL中已有显著的进展,提升模型可解释性仍面临着多个挑战.

(1)高维状态空间和复杂环境. 随着系统规模和任务复杂性的增加,高维状态空间和复杂的智能体交互关系使得决策过程的可解释性变得更为困难. 现有的解释方法往往难以同时处理多智能体之间的交互和状态的动态变化.

(2)多智能体协作的透明性. 在多智能体系统中,智能体不仅需要考虑自身的策略,还需要在合作和博弈中考虑其他智能体的行为. 如何解释多个智能体之间的复杂协作关系,特别是在协作策略或博弈决策中的因果链条,仍然是一个挑战.

(3)因果推理的集成问题. 虽然因果推理可以提高模型的可解释性,但如何有效地将因果推理与MARL算法结合,特别是在处理多个智能体和动态环境时,仍然是一个研究难点. 因果图的构建和推理过程需要精确建模,如何在实时任务中高效地计算因果关系是未来的挑战之一.

(4)可解释性与性能的平衡. 增强可解释性通常意味着需要增加模型的复杂性,这可能会影响系统的性能. 如何在保证高效性和透明度之间找到合理的平衡点,特别是在大规模多智能体系统中,是一个重要的研究课题.

(5)动态任务与环境适应性. 在动态任务和环境变

化中,智能体的策略可能会频繁调整,如何实时提供决策过程的可解释性,并确保新的任务能够快速适应,同时保持高效的可解释性,依然面临很多挑战.

4.4 大规模多智能体系统

在MARL的实际应用中,大规模多智能体系统的训练与部署是面临的重要挑战.随着智能体数量的增加和任务复杂度的提升,传统MARL方法在处理大规模系统时往往会遭遇性能瓶颈,如计算资源消耗过大、通信负担过重、智能体间协作效率低下等问题.因此,如何设计高效的表征机制和合理的模型表达方式,成为提升MARL在大规模异构系统中应用的关键.本节从任务表征和模型表达两个方面综述大规模多智能体系统下的MARL算法,从输入与计算两个维度探讨大规模多智能体系统下的计算机制.

4.4.1 任务表征

在大规模MARL场景中,任务表征能力的提升被认为是缓解训练难度、提升效率与协同性能的关键.近年来,研究者们围绕交通信号控制、模糊逻辑建模、专家知识融合等方向,提出了一系列面向可扩展性的任务表征方法.

文献[110]针对大规模交通信号控制问题,提出了一种基于多智能体A2C的分布式MARL方法,从全局任务分配的角度出发,有效划分局部智能体的控制范围,缓解了联合动作空间维度爆炸与部分可观察性所带来的挑战.实验表明,该方法在合成交通网与真实摩纳哥交通网络中,在优化性、稳健性与样本效率方面均优于独立A2C和IQL方法,展现出在大规模智能交通系统中的应用潜力.进一步地,文献[111]提出了Co-DQL方法,结合独立双Q学习与均值场理论,有效简化了智能体间的交互建模,并通过引入奖励分配机制与局部状态共享策略,提升了训练的稳定性和鲁棒性,体现出良好的协同控制能力.

在任务结构抽象与知识引导方面,文献[112]提出了一种融合人类专家知识的可扩展MARL方法,采用模糊逻辑引导智能体自主吸收非最优专家经验,同时引入基于图的组控制器以提升协同效率与适应能力.实验表明,该方法在应对维度灾难与稀疏奖励等挑战中具有良好的扩展性,适用于交通、能源和机器人系统等复杂环境.文献[113]则进一步提出一种基于模糊逻辑的抽象智能体方法,通过用少量抽象智能体替代众多实体智能体参与训练,并通过模糊策略映射,显著降低计算与空间复杂度,同时隐式捕捉智能体间的耦合关系,在完全可观察与部分可观察的环境下均表现出优异性能,适用于资源受限的实际场景,如智能交通和无人机集群系统.

此外,文献[114]提出的GPLight方法关注交叉口智能体之间的相似性挖掘与参数共享.通过将相似交叉口分组并共享策略网络,GPLight在降低系统复杂度的同时保持策略精度,并引入互信息损失与聚集损失来增强组内稳定性与组间可分离性.实验结果表明,GPLight在多个交通数据集上均优于现有方法(如MP-Light与CoLight),在调度效率、稳定性和可扩展性方面展现出显著优势.

综上所述,从全局任务分配到抽象建模与知识引导,现有研究通过改进任务表征方式,有效推动了大规模MARL系统在实际应用场景中的可扩展性与训练效率,但如何在更加异构、动态的任务环境中进一步提升策略泛化能力与协同自适应性,仍是值得深入探索的方向.

4.4.2 模型表达

在大规模MARL中,优异的模型表达能力对捕获关键特征、提升训练效率和系统扩展性具有重要意义.近期研究围绕注意力机制、GNN、策略优化框架等方向提出了多种创新性方法.

例如,文献[115]提出GAT-MF算法,通过将多智能体间的交互建模为与加权均值场的交互,引入图注意力机制动态捕捉交互强度,在降低计算复杂度的同时提升策略学习的准确性与效率.实验结果显示,该方法在复杂任务中实现了更高的训练效率与更低的GPU内存消耗,特别适用于交通控制与能源管理等大规模协同任务中.文献[116]提出的经典MAAC算法在CTDE框架下设计了基于注意力的集中式Critic网络,为每个智能体动态筛选关键信息,显著提升了训练效率与系统可扩展性,在合作、竞争及混合型任务中均表现出色.

为了应对大规模系统中的通信与信息共享瓶颈,研究者们也提出了去中心化策略优化与局部建模框架.文献[117]提出一种基于模型的去中心化策略优化框架,通过局部模型学习与网络化MDP结构,在缺乏全局信息的情况下实现高效策略更新,适用于交通、电力、疫情防控等复杂场景.文献[118]则采用GNN构建多智能体调度机制,使智能体仅通过本地信息实现去中心化学习与决策,显著降低通信和计算开销,在大规模分布式系统中展现出卓越的调度效率和扩展性.

在MAAC算法基础上,研究者也提出了多种具备领域适应性与系统性扩展能力的衍生方法.文献[119]将MAAC应用于大规模P2P能源管理,关注隐私保护与资源限制问题,在实际住宅场景中显著降低了能耗与峰值负载.文献[120]提出ARMAAC算法,结合递归MAAC与多头注意力机制,有效应对异构边缘计算系统中的资源调度与卸载难题,提升了任务完成率与系统稳定性.

此外,为推动大规模场景下MARL方法的可复现研究与基准评估,文献[121]开发了IMP-MARL环境套件,用于模拟如海上风电场结构维护等复杂任务中的多智能体协作问题. 实验验证表明,CTDE方法(如QMIX、QVMix、QPLEX)在扩展性方面优于集中式与完全去中心化方法. 研究同时指出,在大规模系统中实现智能体间协调仍面临显著挑战,合作机制与可扩展性将是未来研究的重要方向.

综上,当前工作从模型表达、图结构建模、去中心化策略优化等角度出发,显著提升了大规模MARL系统的收敛效率与精度. 但在面对更复杂、动态和异构的现实系统时,如何进一步提升跨智能体的协同效率与跨任务的策略迁移能力,并解决大规模系统中智能体协作和信息共享所带来的如下所示的特有难题,仍然是未来发展的重点.

(1)高维状态空间的处理. 当智能体数量增大时,联合状态空间的名义维度与有效维度同时上升. 名义维度来源于各个体局部观测的堆叠与历史依赖,导致状态张量在时空上高度相关. 有效维度则体现为对关键交互模式、拓扑结构与长期依赖的内在需求. 随着维度的提高,状态空间中的数据分布变得更加稀疏,稀有但关键的协作数据更难以充分覆盖. 同时,部分可观测性与延迟奖励反馈使得对真实系统状态的重建更具挑战. 此外,过高的输入维度还会放大模型表达与优化难度,使表征冗余、泛化脆弱与估计方差升高相互叠加,最终表现为样本效率降低. 因此,如何在不破坏任务结构与对称性的前提下,识别并保留与协作决策最相关的低维“信息子空间”,成为规模化场景下表征层面的巨大挑战之一.

(2)智能体间的有效协作. 协作的本质在于多智能体在时空上的一致性与互补性. 随着系统规模扩大,协作的“关系图”迅速变得稠密而动态,谁与谁互动、何时互动、互动强度多大,均受到任务上下文、资源限制与时延噪声的共同影响. 此外,局部最优行为更易引发整体的相互掣肘(如资源争用、优先级冲突与时序错配),从而削弱全局效用. 由于协作周期长,单次决策对远端个体与未来回报的影响更难被准确感知与利用,协作问题因此呈现出非线性放大的特性,即小规模下可容忍的偏差,在大规模下可能以级联效应形式扩散为系统性退化.

(3)可解释性与透明度. 当参与智能体与交互数量增多、策略结构日益复杂时,个体层面、团队层面与系统层面的因果链条会显著拉长. 大量潜在变量(如隐式通信语义、临时角色分工与环境隐状态)使得从“结果”追溯到“决策依据”的路径被遮蔽,解释对象从单个策略函数扩展为层级化的耦合系统. 大规模系统中还常出现角色切换与策略漂移,使既有解释在时间上迅速过期,难以复用. 与此同时,跨主体的协同行为往往依赖分布层面的统计规律而非单次轨迹的确定性逻辑,这进一步提升了策略解释的门槛. 即便模型的总体性能可观,决策过程对人类与监管方而言也可能是不透明的,影响信任建立、错误诊断与责任划分,并增加在安全关键场景中的部署风险.

(4)分布式学习的优化. 在大规模系统中,分布式学习面临信息局部性与目标一致性的双重张力. 一方面,个体只能基于局部观测与有线通信进行更新,数据分布呈现出强烈的非独立同分布特征,且信息在拓扑上传播存在时延与丢失. 另一方面,个体改进并不必然对应全局目标的改进,局部最优与全局最优之间可能存在系统性偏差. 随着系统规模的扩大,异步更新、陈旧信息与不同个体学习进度的不匹配会加剧算法训练的不稳定性.

(5)隐私保护和安全性. 大规模多智能体场景通常涉及跨域、跨组织的数据协同,即使不直接共享原始数据,梯度、模型参数与通信元数据也可能成为隐私泄露与敏感属性推断的载体. 在开放或半开放环境中,还需面对中毒、回放、旁路监听与合谋等攻击行为. 与此同时,安全与隐私约束会反过来限制学习所需的信息流,形成可用性与合规性的结构性张力,即一味加强保护可能降低学习质量,一味追求性能则可能突破合规红线. 在此背景下,如何在复杂环境下明确定义可共享边界、追踪信息流并保障系统在对抗性扰动下的韧性,成为多智能体规模化部署时不可回避的核心问题.

4.5　多智能体安全强化学习

随着MARL技术在智能交通、自动驾驶、机器人协作等领域的应用,如何确保系统在执行任务时保持安全性成了一个至关重要的问题. 多智能体安全强化学习作为一种新兴的研究领域,旨在使多个智能体能够在复杂环境中协作或竞争,同时确保系统不受外部攻击、恶意行为或意外因素的影响. 在多智能体系统中,安全性不仅是指智能体本身的稳定性,还涉及智能体间的协调安全和对抗性安全. 本节将结合详细案例介绍多智能体安全强化学习的研究现状、关键技术和实际应用.

文献[122]提出了assured MARL方法,引入定量验证技术以在强化学习过程中提供安全性和性能的正式保障. 该方法通过形式化模型检查构建抽象MDP,并在生成安全策略后再应用MARL学习,有效确保智能体在复杂环境中满足安全约束. 实验表明,该方法在异构系统与大规模场景中相较于传统RL更具安全性与任务效率,展现出极大的应用潜力. 文献[123]进一步将安全问题建模为约束马尔可夫博弈,提出一种面向多

机器人控制的安全MARL方法,结合多智能体信任区学习与策略优化,确保策略更新过程中奖励提升与安全约束满足的单调性. 作者还设计了两种基于策略梯度的安全算法,并构建了3个新基准环境(Safe MAMuJoCo、Safe MARobosuite、Safe MAIG),实验结果验证了其在安全性与回报之间的优越平衡. 在自动驾驶场景中,文献[124]提出了一种用于联网自动驾驶汽车行为规划的安全MARL方法,通过基于信息共享的多智能体框架提升决策效率与系统安全性. 该方法引入截断Q函数与安全动作映射机制,在降低计算复杂度的同时提供了可验证的安全保障. CARLA实验表明,该方法在不同交通密度下显著提升了效率与驾驶舒适性,并有效规避不安全动作,展现了在智能车联网中的广泛应用潜力. 为进一步提升高风险环境下的训练安全性,文献[125]提出了DS-MARL方法,通过安全度量与备份策略引导智能体行为,设计"动态盾牌"机制实现训练与执行阶段的不安全动作干预,提升策略收敛性与系统鲁棒性. 在去中心化控制场景中,文献[126]提出了Safe Dec-PG方法,将安全MARL问题建模为网络化约束MDP,结合拉格朗日乘子法与对偶优化策略,在仅使用局部信息的前提下实现策略更新与系统级协调. 该方法引入梯度跟踪与动态调整机制,保障了无中心控制条件下的协同效率,并提供了理论收敛性保证. 最后,文献[127]针对多机器人协作中的团队安全问题,提出了基于软约束策略优化的安全MARL框架. 该方法形式化建模为受约束的马尔可夫博弈,并提出SM-TRPO与SM-PPO两种策略优化算法,以实现每次迭代中奖励的单调改进与安全约束的同时满足,降低了传统方法的计算开销.

总结来看,近年来,研究者们对安全MARL的研究聚焦于提升智能体在协作和竞争中的安全性,确保系统在执行任务时不会受到恶意行为或外部攻击的影响. 主要的研究方向包括通过引入量化验证、软约束策略优化和去中心化策略梯度优化等方法,结合安全度量和备份策略,确保在动态环境下,智能体不仅最大化奖励,还能满足安全约束. 研究中,采用的技术包括截断Q函数、控制障碍函数和动态保护机制等,显著提升了安全性和系统效率. 未来,随着技术的进步,研究可以进一步优化算法的实时性和可扩展性,并通过跨学科融合(如机器人控制、智能交通等领域)推动安全MARL在更复杂环境中的应用.

4.6 本章小结与思考

本章探讨了当前MARL领域的前沿突破,尤其是LLM与MARL结合、Meta-RL、MARL的可解释性、大规模多智能体系统与安全MARL等技术的创新应用. 这些技术虽然在提升系统性能、加速学习过程以及增强智能体协作能力方面展现了巨大潜力,但也面临着诸多挑战.

LLM与MARL的结合,主要通过自然语言的引导和任务规划来加速智能体的学习,并提高协作效率. 通过自然语言的引导,LLM能够在多智能体系统中提供更高层次的任务规划和策略生成,这使得智能体能够更灵活地调整策略,尤其是在复杂的任务和动态环境中. 然而,如何解决指令歧义、推理延迟和算力预算限制,仍是该方法面临的难题. 未来研究可以聚焦于如何提升LLM生成策略的准确性和一致性,以确保在复杂和动态环境中的可靠应用.

Meta-RL为MARL系统带来了跨任务适应性和快速学习能力,使智能体能够在新任务中通过少量样本快速适应,但在面临任务分布不匹配和计算资源限制时,性能可能下降. Meta-RL的关键优势在于让智能体通过在多个任务中积累的经验,能够快速在新任务中进行自我调整,显著提高了样本效率和任务迁移能力. 未来的工作需要关注如何优化任务相似性度量,并在高维和大规模任务中提高样本效率,同时提升任务迁移的稳定性和可靠性.

MARL的可解释性研究,尤其是在因果推理和策略可视化方面,正在推动模型从黑箱向白盒转变,增强用户对系统的信任. 通过引入更加透明的决策过程,MARL的可解释性研究有助于解释智能体如何根据不同的环境信息作出决策,尤其是在高风险领域(如自动驾驶、金融等)的应用中,提升了用户的信任度. 然而,随着任务复杂度的增加,如何在保证可解释性的同时提高模型性能,仍是一个值得进一步探索的问题.

大规模多智能体系统的研究则集中在如何提升计算效率和智能体间的协作能力上. 随着智能体数量的增加,传统的MARL方法面临着计算资源和通信负担的挑战. 当前的研究已提出多种解决方案,如利用GNN和注意力机制来提高计算效率,并且通过局部建模和去中心化的策略优化来减少通信负担. 尽管现有方法在交通、资源调度等领域取得了显著进展,但如何在动态、异构环境中实现更高效的策略迁移和协作,依然是未来的挑战.

随着安全问题在多智能体系统中日益重要,如何在训练和执行阶段实现智能体行为的安全保障,成为确保系统稳定性和高效性的关键. 安全MARL的研究已经开始聚焦于在多智能体系统中嵌入安全机制,通过引入动态保护机制和安全约束,确保智能体在复杂环境下的可靠性. 未来的研究应进一步探索如何在实际部署中结合安全度量和备份策略,确保多智能体系统在高风险环境中稳定运行.

展望未来,随着技术的不断进步,MARL的研究将

朝着更加智能化和自适应的方向发展. 如何让多智能体系统能够在高度不确定和动态的环境中进行快速决策与高效协作,将是一个重要的研究课题. 此外,随着跨学科技术的不断融合,例如与量子计算、大规模并行计算以及5G/6G网络的结合,MARL将在实时决策、智能制造、自动驾驶等领域展现出更大的应用潜力. 未来的研究需要在高效学习算法、跨任务迁移能力和安全可靠性之间找到平衡,并推动MARL技术从理论研究向实际应用转化.

5　MARL的应用部署

随着MARL技术的快速发展,其在各种实际应用中的潜力逐渐显现,特别是在现实多智能体系统中. 传统的MARL方法通常依赖于仿真环境进行训练与测试,但将这些技术从仿真环境成功迁移到现实部署中,仍然面临着一系列挑战. 从仿真到实际应用的过程不仅涉及算法的优化,还需要考虑系统规模、计算资源、通信延迟、隐私保护等现实因素. 因此,如何有效地将MARL技术应用于复杂的多智能体系统中,尤其是要求高效协作、实时响应的环境,成为当前研究的热点之一. 本章将重点探讨MARL在复杂多智能体系统中的应用场景,分析从仿真环境到现实部署过程中所面临的技术挑战与创新解决方案,具体包括多智能体协作、竞争型博弈的应用场景,并进一步探讨在这些实际应用中,如何应对通信延迟、能效、异构系统适配等工程部署中的难题.

5.1　协作应用中的MARL

在一些大规模的协作型任务中,MARL的核心优势体现在其能够高效地处理多个智能体的协同合作问题. 本节以无人机集群应用和智能电网调度领域为例,介绍MARL的协作应用部署.

5.1.1　面向无人机集群应用的MARL

近年来,随着MARL理论与技术的快速发展,研究人员将其广泛应用于无人机集群系统中,以解决编队控制、边缘计算、资源分配、任务调度、目标跟踪等多种协同任务. MARL的协同学习机制为无人机系统在高动态、强耦合、多约束的环境中提供了高效、灵活的决策支持,展现出显著的应用潜力.

文献[128]提出MOIPC-MAAC方法,面向多无人机支持的移动边缘计算系统,解决轨迹规划与任务卸载联合优化问题. 该方法通过引入因果推理通信网络,使无人机在部分可观察环境下能够动态学习通信优先级,提升智能体间协作能力. 同时,采用多目标优化策略,通过广义贝尔曼算子权衡延迟、能耗与任务处理数量,实现了更高效的资源管理与策略迁移. 文献[129]针对多无人机通信网络中的资源管理问题,构建了基于Q-learning的去中心化MARL框架,用于优化用户选择、功率分配与子信道分配. 该方法建模为随机博弈,每个无人机智能体在局部信息基础上独立决策,实现了在不完全信息和高通信成本下的系统性能最大化,优于传统集中式策略. 文献[130]提出Qedgix框架,将GNN与QMIX结合,用于优化多无人机系统中的协作轨迹规划与信息更新管理. GNN用于提取无人机与用户间的关系信息,从而缓解部分观测带来的建模困难. 实验表明,该方法能提升数据采集效率与AoI优化能力,展现出在复杂动态环境中的适应性与稳定性. 文献[131]提出的SiGNN利用无人机编队中存在的空间对称性,引导GNN结构训练,从而提升MARL策略学习效率与样本利用率. 通过在策略网络中引入对称性建模,SiGNN在连续动作空间和部分可观测场景中显著降低了训练开销,适用于大规模协同覆盖任务. 文献[132]聚焦于无人机辅助通信系统的安全性问题,基于MADDPG算法设计了多无人机协同轨迹与功率控制策略,以对抗地面窃听干扰. 该方法通过引入连续动作注意力机制提升了学习效率,实验证明其在高风险通信场景中具备更强的安全性与系统容量. 文献[133]引入对抗性领域随机化与优先经验回放机制,提升了双无人机在复杂任务环境中的适应与迁移能力. 该方法不仅增强了对训练环境多样性的模拟,还显著改善了收敛速度与策略稳定性,在现实运输任务中验证了其实用性与鲁棒性. 文献[134]提出了一种能效优化的MARL编队方法,面向多无人机协同目标跟踪任务. 通过结合动力消耗模型与实时状态信息,该方法实现了策略协作与能耗最小化的统一,有效延长了系统续航时间并提高了任务完成率. 文献[135]构建了基于Transformer的T-MARL方法,解决了大规模多无人机覆盖任务中输入维度变化带来的可扩展性问题. Transformer的注意力机制使系统在面对大规模交互时仍具良好性能,并通过参数共享与预训练技术,提升了模型稳定性与训练效率.

综上所述,基于MARL的多无人机系统研究在多个关键任务场景中取得了重要进展. MARL通过分布式学习、策略协同与动态优化机制,增强了无人机集群在任务分配、资源协调、轨迹生成、通信优化与目标感知等方面的自主决策能力. 未来,随着模型架构(如GNN、Transformer)与推理机制(如因果建模、对抗扰动)的不断引入,MARL在大规模、异构、多目标无人系统中的应用前景将更加广阔,尤其在智能交通、应急响应、军事侦察与空中物流等场景中具备重要的工程应用价值.

5.1.2　面向智能电网调度的MARL

随着智能电网的迅速发展,如何实现高效、可靠的

电力调度与管理成为当前研究的重要方向. 传统电网主要依赖集中式控制系统,但面对动态、复杂且高度不确定的环境,其在应对分布式资源接入、电压波动及系统安全性方面表现出明显局限. 典型的MARL作为一种去中心化的智能优化方法,逐渐成为智能电网调度的关键技术支撑.

文献[136]提出了Power Gridworld,一个专为电力系统设计的开源MARL仿真框架,集成电力流解算器以真实模拟电网物理特性,兼容主流RL训练平台,如RLLib和OpenAI Gym. 该平台支持用户快速搭建多智能体任务环境,解决了电力系统中缺乏标准化测试环境的问题,有助于提升协同策略在实际电网中的可部署性.

在电压调节方向,文献[137]和文献[138]分别将配电网络中的电压控制问题建模为Dec-POMDP,利用MARL方法优化分布式光伏逆变器的无功功率响应,以稳定电压水平. 在此基础上,文献[139]进一步提出EA-MAAC算法,融合注意力机制与双时间尺度控制策略,实现对有源与无源设备的分层调控,显著提升了系统的响应速度与调节效率.

文献[140]针对工业智能电网中的能源优化问题,提出将MARL应用于能源供需调节、市场交易及储能管理. 该方法通过多代理系统协调运行,显著降低了短期市场波动带来的能耗成本,在动态生产环境中展现出更强的应对能力与经济性.

在系统协调控制方面,文献[141]提出了CEQ算法,实现多区域电网的联合负荷频率控制. 该算法通过时变平衡因子动态识别最优协作策略,强化区域间的信息共享和战略联动,提升了系统的长期运行性能. 文献[142]将MADDPG应用于高渗透率光伏系统下的电压调节,构建多智能体协同控制架构,使每个逆变器可自主进行电压调节决策,增强系统的可扩展性与稳定性. 文献[143]进一步结合主动和无功功率联合优化,提出去中心化协调机制,有效应对不同负荷场景下的电压扰动问题.

为提高模型的容错性与通信效率,文献[144]提出C-MARL框架,通过代理间的价值函数一致性学习,实现分布式电压控制. 该方法无需精确网络建模,具备良好的扩展性与通信鲁棒性,在面对代理失效或链路中断时仍能维持调控效果.

综上所述,MARL在智能电网领域已展现出广泛的应用前景. 无论是电压控制、频率调节、能量交易还是资源协调,MARL均能通过去中心化学习机制实现高效、稳健的系统优化. 通过引入共识机制、状态压缩、注意力机制及多尺度控制,现有研究不断突破样本效率、适应性与协作性等方面的瓶颈,逐步推动MARL向真实智能电网系统的落地部署迈进.

5.2 竞争应用中的MARL

在交通管理、自动驾驶、电力市场与博弈系统等诸多实际场景中,竞争性问题普遍存在,构成多智能体系统中的核心挑战之一. 与合作性任务不同,竞争性任务中各智能体拥有独立目标,强化了对策略鲁棒性、适应性与泛化能力的需求. MARL凭借其策略优化能力,已成为解决复杂竞争环境中智能体博弈行为的重要手段.

为应对竞争环境中的泛化问题,文献[145]提出GMARL,通过引入差分隐私机制对观察数据进行拉普拉斯扰动,从而增强智能体在动态与不确定环境下的泛化能力. 该方法有效降低了传统MARL模型对环境扰动的敏感性,使智能体在策略迁移时保持稳定性.

针对零和博弈中的策略优化,文献[146]提出了独立策略梯度算法,采用双时间尺度更新规则实现策略收敛. 该方法无需对手策略协调即可达到纳什均衡,是首个在零和随机博弈中提供有限样本收敛保证的独立学习方法,提升了MARL在对抗性场景中的适用性与理论可信度.

在竞争性资源定价问题上,文献[147]研究了共享自动驾驶汽车与人力驾驶平台的竞争定价策略. 基于MADDPG框架,提出空间—时间动态定价模型,优化在部分可观测环境中的定价行为. 结果显示,共享平台即使车队规模较小,也可通过策略优化获取更高利润,验证了MARL在经济博弈建模中的潜力.

文献[148]将MARL与迁移学习结合,提出了一种智能竞标策略,应用于电力市场中发电商和购电方之间的策略博弈. 其中,发电方采用Q-learning自主优化竞价策略,买方则借助迁移学习提升市场需求预测精度,显著提升了博弈策略的学习效率与市场适应性.

在经济系统建模方面,文献[149]提出了基于竞争均衡的MARL算法,融合交换经济理论与RL学习框架,使多个自利型智能体在动态系统中达成竞争均衡. 通过在线/离线算法实现对不确定效用函数的鲁棒学习,为多智能体系统中的博弈建模提供了理论支撑.

在强化学习方法研究方面,文献[150]提出了基于自对弈与置信区间的竞争性MARL算法. 该方法通过构建值函数的上下置信界估计,引导策略探索与稳定收敛. 同时,创新性地使用“先探索后利用”的机制,使得智能体在完全对抗环境中也能实现自主高效训练,避免对外部专家的依赖.

综上所述,在复杂的竞争性环境中,MARL结合策略泛化、对抗博弈建模、自对弈机制及迁移学习等技术,已在理论与实践两个层面取得显著突破. 这些方法不仅扩展了MARL在实际系统(如交通、经济、电力)中

的应用边界,也为多智能体系统在对抗与竞合中的策略优化提供了坚实基础.

5.3 本章小结与思考

本章重点讨论了MARL在现实多智能体系统中的应用部署,特别是在协作性和竞争性任务中的实际挑战. 随着技术的不断发展,MARL在无人机集群、智能电网调度、自动驾驶等领域展现了巨大潜力. 然而,从仿真环境到现实部署的过渡,面临着诸多工程难题,如通信延迟、能效管理和异构系统的适配等. 在具体应用中,这些问题都需要得到有效解决.

在协作应用方面,MARL的分布式学习和动态优化机制,使得无人机集群和智能电网等领域能够实现高效的资源调度和协同决策. 在具体应用中,采用分布式学习和多智能体协作,使得系统能够在复杂动态环境中快速响应并优化任务分配. 然而,如何进一步提升协作效率并减少通信延迟,依然是当前研究的重点. 未来,随着更多先进算法(如GNN和Transformer)的引入,MARL将在智能交通和空中物流等场景中展现出更大的应用潜力.

在竞争应用方面,MARL的博弈理论和策略优化能力,使其在交通管理、自动驾驶、电力市场等领域发挥着重要作用. 通过引入迁移学习和自对弈机制,这些方法不仅提升了模型的泛化能力,还能应对复杂的对抗博弈环境. 尽管如此,如何平衡鲁棒性与策略收敛性,以及如何优化竞争博弈中的资源定价等,仍然是MARL应用中的难点. 未来的研究可以关注如何通过更高效的博弈建模和算法优化,提高系统在不确定环境中的表现.

未来,随着MARL在多智能体系统中的广泛应用,其部署将更加注重实际工程中的约束条件. 随着通信预算、能耗和实时响应等实际要求的增加,如何在这些约束下实现MARL算法的高效运行,将成为研究的重点. 此外,随着跨领域协同和多组织合作的需求增加,如何通过标准化接口和协议,确保系统的稳定性、可扩展性和可审计性,依然是未来应用成功的关键. 总的来说,未来的MARL不仅要学得好,更要跑得稳,确保在复杂、多变的现实环境中稳定运行并具有高度适应性.

6 结论与展望

6.1 未来研究趋势

随着MARL技术的不断进步和应用领域的扩展,未来的研究将集中在多个关键方向,从可扩展性、可解释性、大模型结合等方面进行创新. 以下是未来MARL研究的7个主要趋势,旨在推动该领域在更多复杂环境和实际应用中实现广泛应用.

6.1.1 可扩展性:从数量增长到复杂度不升

当前MARL的可扩展性研究仍面临巨大挑战,即随着智能体数量和状态与动作空间维度的增加,模型表达、学习效率和硬件算力均受到严峻考验. 现有方法如CTDE、均值场近似等在一定程度上缓解了非平稳性和维度灾难,但算法在大规模复杂场景下仍存在收敛慢、性能退化的瓶颈. 为取得创新突破,可考虑引入分层强化学习框架,将大规模任务分解为若干层次子任务以降低决策复杂度. 同时,结合GNN对智能体交互关系建模,利用交互拓扑的稀疏性来提升学习效率和可扩展上限. 另外,借助并行计算和联邦学习等分布式训练范式,可在保持收敛性的前提下支持数百甚至上千个智能体的协同训练,实现规模扩展的跨越式提升. 综上,从算法架构和工程机制两方面协同发力,有望解决MARL在超大规模环境中的可扩展性难题.

6.1.2 可解释性:提升智能体决策过程的透明度

深度MARL策略往往呈现内部难以解释的“黑箱”结构,使人类难以理解、信任和监督智能体决策. 目前,一些研究尝试通过值函数分解、注意力机制等手段提高多智能体决策的透明度,例如用注意力权重度量各智能体贡献或利用可视化工具诊断策略模式,但多智能体协作或对抗过程中策略涌现的机理尚缺乏清晰的理论框架. 现存瓶颈在于如何解释智能体间复杂交互下的信用分配和决策因果? 如何提取高维策略网络中的人类可读知识? 为实现突破,需从引入可解释模型和增强事后分析两方面入手. 一方面,可在设计阶段优先采用具备内在可解释性的方法(例如将决策规则提取为决策树或符号表示等)来构建智能体或对策略进行符号提炼,以生成可解释的策略描述;另一方面,融合因果推理的方法有望揭示智能体决策背后的因果链条,构建智能体交互的因果图谱. 此外,引入LLM生成自然语言的决策理由,作为智能体的“自我解释”模块,亦是增强可解释性的大胆尝试. 这些措施将有助于提升多智能体系统的透明度和可审查性,为关键领域的部署奠定基础.

6.1.3 LLM结合:深度学习与多智能体强化学习的深度融合

LLM的崛起为MARL带来了新的机遇与挑战. 现有研究已开始探索将LLM等大语言模型融入多智能体学习框架. 让知识丰富的LLM为智能体提供任务指导,可显著提高探索效率和学习速度. 然而,LLM与MARL的深度融合尚存在理论空白. LLM虽然在单智能体决策中展现出强大能力,但扩展到多智能体环境并不容易,因为传统LLM缺乏针对智能体协调与通信的机制. 同时,大模型推理开销巨大,如何在保证实时交互的前提下发挥其全局知识也是一大瓶颈. 针对这

些问题,未来可探索以下创新路径:

(1)研制面向决策的基础模型,将多智能体交互机制融入预训练过程,使模型天然具备协同决策能力.

(2)采用分层决策架构. 高层由大模型进行策略规划和意图理解,底层由强化学习智能体执行具体动作,实现符号推理与数值决策的结合.

(3)利用大模型进行环境建模和奖励塑形,例如通过提示引导LLM动态生成子目标或评估策略表现,从而提升样本效率和泛化能力.

综上所述,大模型与MARL的结合需要在算法上设计出兼顾知识引入与交互协调的新范式,充分发挥大模型的知识优势来突破当前MARL的性能天花板.

6.1.4 多领域技术的深度融合:强化学习与其他学科的结合

MARL的发展有赖于跨学科的深度融合. 目前,控制论和博弈论等领域的方法已部分融入MARL,例如安全约束强化学习、鲁棒训练和元学习等方向,初步证明了跨领域融合在提升算法可靠性和适应性方面的价值. 然而,现有研究主要停留在借鉴表层概念,尚未形成系统的跨学科理论框架,难以应对现实系统中复杂多变的问题. 未来的突破路径在于构建统一的跨学科研究范式. 一方面,在理论层面引入复杂系统科学观点,将MARL视作动态博弈系统,将经济学中的机制设计、演化博弈理论等引入奖励设计和策略涌现分析. 另一方面,在应用层面融合机器人学、控制工程等知识,在算法中显式加入物理约束、安全规则等领域先验,以确保智能体决策既高效又符合领域规范. 例如,可借鉴经济学的拍卖机制设计团队合作策略,或利用生物学中的群体智能启发式提高协作效率. 通过建立跨领域的协同研究团队,让博弈论、认知科学、控制工程等专家共同参与MARL算法设计,可产出兼具理论严谨性和工程实用性的创新方案,从而突破单一学科局限,推动多智能体智能向更高层次发展.

6.1.5 实时性与延迟优化:适应高动态环境的强化学习

实时决策能力是多智能体系统走向实际应用的关键需求之一. 例如,在自动驾驶、智能电网等场景中,智能体必须在毫秒级别响应环境变化. 然而,当前MARL算法往往需要大量迭代训练,难以及时适应环境动态. 即使训练得到的策略,在部署时也可能因计算延迟或通信迟滞而无法实时响应. 通信延迟会削弱智能体间协同决策的时效性,这突出体现了MARL在实时性方面的技术瓶颈. 现有一些工作尝试通过异步更新、增量学习等提高在线适应能力,并利用模型预测控制等方法在执行阶段进行快速滚动优化,以弥补纯强化学习决策的迟滞. 但是,实现真正的实时MARL仍需进一步创新. 可能的突破路径包括以下3种:(1)发展“随时可用”的在线学习算法,让智能体在运行过程中持续学习,并能在任意时间输出当前最优策略近似,以应对快速变化的环境;(2)引入事件触发机制,当环境出现突发变化时,及时激活学习更新,从而减少不必要的计算开销,保障关键时刻的响应速度;(3)软硬件协同优化,在算法层面简化网络结构、减少决策所需计算量,在系统工程上采用专用加速芯片和高速通信架构,降低决策延迟.

通过以上措施的综合作用,多智能体系统有望在复杂动态环境中实现毫秒级的自主协同响应,满足实际应用对实时性的严格要求.

6.1.6 异构平台和设备适配:智能体系统的多平台部署

在开放动态的多智能体环境中,智能体往往具有异构性. 不同的智能体可能拥有不同的传感器、动作空间或目标偏好,甚至在运行过程中会有新的智能体加入或现有智能体退出. 这对传统假设同构和固定智能体集合的MARL提出了严峻挑战. 简单的参数共享策略会导致所有智能体策略同质化,不利于发挥异构智能体各自的优势. 自适应的部分参数共享机制通过引入可学习的网络,使每个智能体在共享基础能力的同时保留差异化策略,以兼顾协作效率和策略多样性. 尽管这类方法在一定程度上提升了异构环境下的性能,但仍存在技术瓶颈. 当环境中出现全新类型的智能体时,如何迅速适配? 不同智能体数量变化时,如何保持既有策略的鲁棒性? 未来的创新方向应侧重于提高策略的泛化和适应能力.

(1)加强元学习和迁移学习路线. 预先训练一个能够适配不同智能体类型的加强元策略,当引入新智能体时,仅需少量交互便可快速调整策略.

(2)策略模块化路线. 将策略拆解为公共模块和特定模块,公共模块学习通用协作原理,特定模块针对每类智能体特有能力进行微调,从而实现“即插即用”的策略扩展.

(3)开放代理体系结构. 设计智能体加入和离开的协议和算法,如教师—学生框架,让新智能体通过已有智能体学习策略,或通过在线微调使系统对组分变化具备弹性适应能力. 通过上述机制,多智能体系统将能够更从容地应对异构环境和动态组成,在开放复杂的现实世界中保持稳定的合作状态与性能.

6.1.7 安全性与隐私保护:增强MARL系统的可靠性和保护机制

多智能体系统在安全敏感和隐私敏感的场景(如无人驾驶编队、分布式能源管理、协同医疗决策等)中应用日益广泛,对安全与隐私的要求愈发严格. 传统MARL主要关注奖励值优化,往往忽视了策略的安全约

束和训练过程中的信息泄露风险. 这导致智能体可能会习得人类不可接受的危险行为,或在通信中泄露敏感信息. 在实际系统中,各参与智能体往往出于安全或隐私考虑,不愿直接共享各自领域的局部信息,这给协作带来了额外困难. 展望未来,一条重要的创新路径是将安全与隐私"内生化"到MARL的框架设计中,即从算法设计初始阶段就嵌入安全约束和隐私保护机制. 例如,采用联合训练但不共享原始数据的联邦强化学习,结合安全多方计算或同态加密,确保各智能体在不暴露私有信息的情况下协同更新策略. 又如,引入形式化验证工具,对多智能体策略进行安全属性验证,在训练过程中实时排除潜在的不安全决策. 更进一步的方向包括建立信任度量和异常检测机制,及时发现并惩罚恶意智能体行为,以及制定多智能体系统的隐私评估标准和合规框架. 通过技术和制度的双管齐下,未来的MARL系统将在保证协作效率的同时最大限度地满足安全可靠和隐私保护要求,为其在关键领域的落地保驾护航.

6.2　总结

本文系统梳理了近年来MARL在基础理论、建模框架、前沿热点等方面的研究工作,旨在提供一种多维度的MARL综述分析. 具体来说,本文首先介绍了MARL的理论建模与关键概念,并从策略梯度、值函数、模型驱动3个维度分析了MARL的核心技术. 其次,本文从LLM、Meta-RL、大规模多智能体系统、可解释性、多智能体安全强化学习这5个方向重点分析了MARL的前沿研究工作,并总结了其未来面临的挑战. 然后,本文从MARL现实部署的角度出发,总结了MARL在协作与竞争应用中的研究工作,并从通信延迟、能耗、异构接口适配等多个角度分析了MARL的工程部署挑战. 最后,本文分析列出了MARL的未来研究方向.

参考文献

[1] TANG C, ABBATEMATTEO B, HU J H, et al. Deep reinforcement learning for robotics: A survey of real-world successes[C]//Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. New York: ACM, 2025: 28694-28698.

[2] MILANI S, TOPIN N, VELOSO M, et al. Explainable reinforcement learning: A survey and comparative review[J]. ACM Computing Surveys, 2024, 56(7): 3616864.

[3] TANG Y L, SUN J, WANG H, et al. A method of network attack-defense game and collaborative defense decision-making based on hierarchical multi-agent reinforcement learning[J]. Computers & Security, 2024, 142: 103871.

[4] SHI H R, LIU G J, ZHANG K W, et al. MARL Sim2real transfer: Merging physical reality with digital virtuality in metaverse[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2023, 53(4): 2107-2117.

[5] YOUN J, PARK J, KIM S, et al. MARL-based access control for grant-free nonorthogonal random access in UDN[J]. IEEE Internet of Things Journal, 2024, 11(17): 28421-28436.

[6] 陈阳, 皮德常, 代成龙, 等. 多无人机协同陆地设施辅助移动边缘计算的系统能耗最小化方法[J]. 电子学报, 2023, 51(4): 984-992.
CHEN Y, PI D C, DAI C L, et al. System energy consumption minimization method for multi-UAVs cooperating with land facilities to assist moving edge calculation[J]. Acta Electronica Sinica, 2023, 51(4): 984-992. (in Chinese)

[7] ZHANG H, CHENG J Y, ZHANG L, et al. H2GNN: Hierarchical-hops graph neural networks for multi-robot exploration in unknown environments[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 3435-3442.

[8] PEY J J J, SAMARAKOON S M B P, MUTHUGALA M A V J, et al. A Decentralized Partially Observable Markov Decision Process for complete coverage onboard multiple shape changing reconfigurable robots[J]. Expert Systems with Applications, 2025, 271: 126565.

[9] ABOUELAZM A, MICHEL J, ZÖLLNER J M. A review of reward functions for reinforcement learning in the context of autonomous driving[C]//2024 IEEE Intelligent Vehicles Symposium. Piscataway: IEEE, 2024: 156-163.

[10] 彭翔, 许华, 蒋磊, 等. 一种基于深度强化学习的动态自适应干扰功率分配方法[J]. 电子学报, 2023, 51(5): 1223-1234.
PENG X, XU H, JIANG L, et al. Dynamic adaptive interference power allocation method based on deep reinforcement learning[J]. Acta Electronica Sinica, 2023, 51(5): 1223-1234. (in Chinese)

[11] SESSA P G, KAMGARPOUR M, KRAUSE A. Efficient model-based multi-agent reinforcement learning via optimistic equilibrium computation[EB/OL]. (2022-07-10)[2025-10-10]. https://arXiv.org/abs/2203.07322.

[12] ESCHMANN J. Reward function design in reinforcement learning[M]//Reinforcement Learning Algorithms: Analysis and Applications. Cham: Springer International Publishing, 2021: 25-33.

[13] TORO ICARTE R, KLASSEN T Q, VALENZANO R, et al. Reward machines: Exploiting reward function structure in

reinforcement learning[J]. Journal of Artificial Intelligence Research, 2022, 73: 173-208.

[14] TABRIZI Y H, UDDIN M N. Multi-agent reinforcement learning-based maximum power point tracking approach to fortify PMSG-based WECSs[J]. IEEE Transactions on Industry Applications, 2024, 60(6): 8077-8087.

[15] RETZLAFF C O, DAS S, WAYLLACE C, et al. Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities[J]. Journal of Artificial Intelligence Research, 2024, 79: 359-415.

[16] 蔡玉, 官铮, 王增文, 等. 基于多智能体深度强化学习的车联网区分业务资源分配算法[J]. 计算机工程与科学, 2024, 46(10): 1757-1764.
CAI Y, GUAN Z, WANG Z W, et al. Multi-agent deep reinforcement learning based resource allocation algorithm for differentiated services in Internet of vehicles[J]. Computer Engineering and Science, 2024, 46(10): 1757-1764. (in Chinese)

[17] YAN Y M, CHOW A H F, HO C P, et al. Reinforcement learning for logistics and supply chain management: Methodologies, state of the art, and future opportunities[J]. Transportation Research Part E: Logistics and Transportation Review, 2022, 162: 102712.

[18] 徐少毅, 杨磊. 基于多智能体深度强化学习的多无人机辅助移动边缘计算轨迹设计[J]. 北京交通大学学报, 2024, 48(5): 1-9.
XU S Y, YANG L. Trajectory design of multi-UAV-assisted moving edge calculation based on multi-agent depth reinforcement learning[J]. Journal of Beijing Jiaotong University, 2024, 48(5): 1-9. (in Chinese)

[19] CHEN D, ZHANG K X, WANG Y Q, et al. Communication-efficient decentralized multi-agent reinforcement learning for cooperative adaptive cruise control[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(10): 6436-6449.

[20] ZHANG K Q, YANG Z R, BAŞAR T. Multi-agent reinforcement learning: A selective overview of theories and algorithms[M]//Handbook of Reinforcement Learning and Control. Cham: Springer International Publishing, 2021: 321-384.

[21] GRONAUER S, DIEPOLD K. Multi-agent deep reinforcement learning: A survey[J]. Artificial Intelligence Review, 2022, 55(2): 895-943.

[22] DU W, DING S F. A survey on multi-agent deep reinforcement learning: From the perspective of challenges and applications[J]. Artificial Intelligence Review, 2021, 54(5): 3215-3238.

[23] HU K, LI M Y, SONG Z Q, et al. A review of research on reinforcement learning algorithms for multi-agents[J]. Neurocomputing, 2024, 599: 128068.

[24] ZHU C X, DASTANI M, WANG S H. A survey of multi-agent deep reinforcement learning with communication[J]. Autonomous Agents and Multi-Agent Systems, 2024, 38(1): 2845-2847.

[25] LIU Z H, ZHANG J Y, SHI E Y, et al. Graph neural network meets multi-agent reinforcement learning: Fundamentals, applications, and future directions[J]. IEEE Wireless Communications, 2024, 31(6): 39-47.

[26] DU P, LI F L, SHAO J L. Multi-agent reinforcement learning clustering algorithm based on silhouette coefficient[J]. Neurocomputing, 2024, 596: 127901.

[27] YU W W, WANG R, HU X H. Learning attentional communication with a common network for multiagent reinforcement learning[J]. Computational Intelligence and Neuroscience, 2023, 2023: 5814420.

[28] HAN H M, JIANG X, LU W D, et al. A multi-agent reinforcement learning approach for massive access in NOMA-URLLC networks[J]. IEEE Transactions on Vehicular Technology, 2023, 72(12): 16799-16804.

[29] CHEN J D, LAN T, JOE-WONG C. RGMComm: Return gap minimization via discrete communications in multi-agent reinforcement learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(16): 17327-17336.

[30] YOUNAS R, RAZA UR REHMAN H M, LEE I, et al. SA-MARL: Novel self-attention-based multi-agent reinforcement learning with stochastic gradient descent[J]. IEEE Access, 2025, 13: 35674-35687.

[31] HAN R X, LI H X, KNOBLOCK E J, et al. Joint velocity and spectrum optimization in urban air transportation system via multi-agent deep reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2023, 72(8): 9770-9782.

[32] LIN H Y, LYU C, HE Y X, et al. Enhancing state representation in multi-agent reinforcement learning for platoon-following models[J]. IEEE Transactions on Vehicular Technology, 2024, 73(8): 12110-12114.

[33] KARPE M, FANG J, MA Z Y, et al. Multi-agent reinforcement learning in a realistic limit order book market simulation[EB/OL]. (2020-06-10) [2025-10-10]. https://arXiv.org/abs/2006.05574.

[34] JI Y X, WANG Y, ZHAO H T, et al. Multi-agent reinforce-

ment learning resources allocation method using dueling double deep Q-network in vehicular networks[J]. IEEE Transactions on Vehicular Technology, 2023, 72(10): 13447-13460.

[35] LI H L, YI P, WEI D X, et al. Seek-and-take games of heterogeneous agent teams with large language model[C]// 2024 China Automation Congress. Piscataway: IEEE, 2025: 7078-7084.

[36] YANG T T, FENG P, GUO Q X, et al. AutoHMA-LLM: Efficient task coordination and execution in heterogeneous multi-agent systems using hybrid large language models[J]. IEEE Transactions on Cognitive Communications and Networking, 2025, 11(2): 987-998.

[37] CHEN C L, WANG Z, WU W H, et al. Meta-DT: Offline meta-RL as conditional sequence modeling with world model disentanglement[C]//Advances in Neural Information Processing Systems 37. Berkeley: USENIX Association, 2024: 44845-44870.

[38] LEE S, CHUNG S Y. Improving generalization in meta-RL with imaginary tasks from latent dynamics mixture[EB/OL]. (2022-01-18)[2025-10-10]. https://arXiv.org/abs/2105.13524.

[39] WANG H, YU Y, JIANG Y. Fully decentralized multiagent communication via causal inference[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(12): 10193-10202.

[40] WANG C, TANG H Z, DING W B. MAMGDT: Enhancing multi-agent systems with multi-game decision transformer[C]//Proceedings of the 30th Annual International Conference on Mobile Computing and Networking. New York: ACM, 2024: 1962-1967.

[41] LI Y H, ZHANG X X, ZENG T Y, et al. Task placement and resource allocation for edge machine learning: A GNN-based multi-agent reinforcement learning paradigm[J]. IEEE Transactions on Parallel and Distributed Systems, 2023, 34(12): 3073-3089.

[42] DAI Y P, LYU L, CHENG N, et al. A survey of graph-based resource management in wireless networks: Part II: Learning approaches[J]. IEEE Transactions on Cognitive Communications and Networking, 2025, 11(4): 2101-2122.

[43] LITTMAN M L. Markov games as a framework for multi-agent reinforcement learning[M]//Machine Learning Proceedings 1994. Amsterdam: Elsevier, 1994: 157-163.

[44] BEYNIER A, CHARPILLET F, SZER D, et al. DEC-MDP/POMDP[M]//Markov Decision Processes in Artificial Intelligence. Hoboken: Wiley, 2013: 277-318.

[45] MA C Y T, YAU D K Y, LOU X, et al. Markov game analysis for attack-defense of power networks under possible misinformation[J]. IEEE Transactions on Power Systems, 2013, 28(2): 1676-1686.

[46] MURPHY K P. A survey of POMDP solution techniques[J]. Environment, 2000, 2(10): 268076619.

[47] DIBANGOYE J S, AMATO C, BUFFET O, et al. Optimally solving dec-POMDPs as continuous-state MDPs[J]. Journal of Artificial Intelligence Research, 2016, 55: 443-497.

[48] KRAEMER L, BANERJEE B. Multi-agent reinforcement learning as a rehearsal for decentralized planning[J]. Neurocomputing, 2016, 190: 82-94.

[49] DENG Y, WANG Z R, CHEN X, et al. Boosting multi-agent reinforcement learning via contextual prompting[J]. Journal of Machine Learning Research, 2023, 24(399): 1-34.

[50] MIAO C Y, CUI Y D, LI H Y, et al. Effective multi-agent deep reinforcement learning control with relative entropy regularization[J]. IEEE Transactions on Automation Science and Engineering, 2025, 22: 3704-3718.

[51] KIM J B, CHOI H B, HAN Y H. Strangeness-driven exploration in multi-agent reinforcement learning[J]. Neural Networks, 2024, 172: 106149.

[52] FENG P, LIANG J K, WANG S Z, et al. Hierarchical consensus-based multi-agent reinforcement learning for multi-robot cooperation tasks[C]//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE, 2024: 642-649.

[53] PEI Y H, REN T, ZHANG Y X, et al. Policy distillation for efficient decentralized execution in multi-agent reinforcement learning[J]. Neurocomputing, 2025, 626: 129617.

[54] GUPTA J K, EGOROV M, KOCHENDERFER M. Cooperative multi-agent control using deep reinforcement learning[C]//Autonomous Agents and Multiagent Systems. Cham: Springer, 2017: 66-83.

[55] ITURRIA-RIVERA P E, CHENIER M, HERSCOVICI B, et al. Channel selection for Wi-Fi 7 multi-link operation via optimistic-weighted VDN and parallel transfer reinforcement learning[C]//2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications. Piscataway: IEEE, 2023: 1-6.

[56] RASHID T, SAMVELYAN M, DE WITT C S, et al. Monotonic value function factorisation for deep multi-agent reinforcement learning[EB/OL]. (2020-08-27) [2025-10-10].

https://arXiv.org/abs/2003.08839.

[57] SON K, KIM D, KANG W J, et al. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning[EB/OL]. (2019-05-14) [2025-10-10]. https://arXiv.org/abs/1905.05408.

[58] WANG J H, REN Z Z, LIU T, et al. QPLEX: Duplex dueling multi-agent Q-learning[EB/OL]. (2021-10-04)[2025-10-10]. https://arXiv.org/abs/2008.01062.

[59] FOERSTER J, FARQUHAR G, AFOURAS T, et al. Counterfactual multi-agent policy gradients[EB/OL]. (2024-12-11)[2025-10-10]. https://arXiv.org/abs/1705.08926.

[60] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[EB/OL]. (2020-03-14)[2025-10-10]. https://arXiv.org/abs/1706.02275.

[61] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of PPO in cooperative, multi-agent games[EB/OL]. (2022-11-04)[2025-10-10]. https://arXiv.org/abs/2103.01955.

[62] WANG T H, DONG H, LESSER V, et al. ROMA: Multi-agent reinforcement learning with emergent roles[EB/OL]. (2020-07-04)[2025-10-10]. https://arXiv.org/abs/2003.08039.

[63] PENG B, RASHID T, DE WITT C A S, et al. FACMAC: Factored multi-agent centralised policy gradients[EB/OL]. (2021-05-07)[2025-10-10]. https://arXiv.org/abs/2003.06709.

[64] KUBA J G, CHEN R Q, WEN M N, et al. Trust region policy optimisation in multi-agent reinforcement learning[EB/OL]. (2022-04-04)[2025-10-10]. https://arXiv.org/abs/2109.11251.

[65] WEN M N, KUBA J G, LIN R J, et al. Multi-agent reinforcement learning is a sequence modeling problem[C]// Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: ACM, 2022: 16509-16521.

[66] TAN M. Multi-agent reinforcement learning: Independent *vs*. cooperative agents[M]//Machine Learning Proceedings 1993. Amsterdam: Elsevier, 1993: 330-337.

[67] STEPANOV E P, SMELIANSKY R L, PLAKUNOV A V, et al. On fair traffic allocation and efficient utilization of network resources based on MARL[J]. Computer Networks, 2024, 250: 110540.

[68] ZHU S C, HAN G J, LIN C. A software-defined MARL-based architecture for AUV cluster network to enable cooperative and smart underwater target tracking[J]. IEEE Wireless Communications, 2024, 31(6): 56-62.

[69] ZHANG K Q, YANG Z R, LIU H, et al. Fully decentralized multi-agent reinforcement learning with networked agents[EB/OL]. (2018-02-27)[2025-10-10]. https://arXiv.org/abs/1802.08757.

[70] KOPPEL A, SINGH BEDI A, GANGULY B, et al. Convergence rates of average-reward multi-agent reinforcement learning via randomized linear programming[C]// 2022 IEEE 61st Conference on Decision and Control. Piscataway: IEEE, 2023: 4545-4552.

[71] WATKINS C J C H, DAYAN P. Q-learning[J]. Machine Learning, 1992, 8(3): 279-292.

[72] SUTTON R S. Generalization in reinforcement learning: Successful examples using sparse coarse coding[C]//Proceedings of the 9th International Conference on Neural Information Processing Systems. New York: ACM, 1995: 1038-1044.

[73] BERTSEKAS D. Multiagent reinforcement learning: Rollout and policy iteration[J]. IEEE/CAA Journal of Automatica Sinica, 2021, 8(2): 249-272.

[74] SILVER D, LEVER G, HEESS N M O, et al. Deterministic policy gradient algorithms[C]//International Conference on Machine Learning. Brookline: JMLR, 2014: 605-619.

[75] MOERLAND T M, BROEKENS J, PLAAT A, et al. Model-based reinforcement learning: A survey[EB/OL]. (2022-03-31)[2025-10-10]. https://arXiv.org/abs/2006.16712.

[76] DU Y L, MA C D, LIU Y C, et al. Scalable model-based policy optimization for decentralized networked systems[C]// 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE, 2022: 9019-9026.

[77] DONG S, XIA Y J, PENG T. Network abnormal traffic detection model based on semi-supervised deep reinforcement learning[J]. IEEE Transactions on Network and Service Management, 2021, 18(4): 4197-4212.

[78] WILLEMSEN D, COPPOLA M, DE CROON G C H E. MAMBPO: Sample-efficient multi-robot reinforcement learning using learned world models[C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). New York: ACM, 2021: 5635-5640.

[79] JANNER M, FU J, ZHANG M, et al. When to trust your model: Model-based policy optimization[EB/OL]. (2021-11-29)[2025-10-10]. https://arXiv.org/abs/1906.08253.

[80] EGOROV V, SHPILMAN A. Scalable multi-agent model-based reinforcement learning[EB/OL]. (2022-05-25)[2025-10-10]. https://arXiv.org/abs/2205.15023.

[81] NAM D, MACVEAN A, HELLENDOORN V, et al. Using

an LLM to help with code understanding[C]//Proceedings of the IEEE/ACM 46th International Conference on Software Engineering. New York: ACM, 2024: 1184-1196.

[82] AN S N, CHEN W Z, LIN Z Q, et al. Make your LLM fully utilize the context[C]//Advances in Neural Information Processing Systems 37. Berkeley: USENIX Association, 2024: 62160-62188.

[83] ALBERTS I L, MERCOLLI L, PYKA T, et al. Large language models (LLM) and ChatGPT: What will the impact on nuclear medicine be?[J]. European Journal of Nuclear Medicine and Molecular Imaging, 2023, 50(6): 1549-1552.

[84] ZHU G B, ZHOU R, JI W K, et al. LAMARL: LLM-aided multi-agent reinforcement learning for cooperative policy generation[J]. IEEE Robotics and Automation Letters, 2025, 10(7): 7476-7483.

[85] YAO T L, XU Y Q, WANG H Y, et al. Multi-agent fuzzy reinforcement learning with LLM for cooperative navigation of endovascular robotics[J]. IEEE Transactions on Fuzzy Systems, 2025. DOI:10.1109/TFUZZ.2025.3585934.

[86] LI Z M, ZHANG R B, WANG Z M, et al. LLM-guided decision-making toolkit for multi-agent reinforcement learning[J]. Neurocomputing, 2025, 638: 130105.

[87] CHALAKI B, LEE K, LEWIS M, et al. Language grounded multi-agent reinforcement learning with human-interpretable communication[EB/OL]. (2024-09-25)[2025-10-10]. https://arXiv.org/pdf/2409.17348.

[88] MORAD S, SHANKAR A, BLUMENKAMP J, et al. Language-conditioned offline RL for multi-robot navigation[C]//2025 IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2025: 14984-14991.

[89] ZHOU L, DENG X F, WANG Z, et al. Semantic information extraction and multi-agent communication optimization based on generative pre-trained transformer[J]. IEEE Transactions on Cognitive Communications and Networking, 2025, 11(2): 725-737.

[90] LOU J B, SHI R Y, LIN Y X, et al. TALKER: A task-activated language model based knowledge-extension reasoning system[J]. IEEE Robotics and Automation Letters, 2025, 10(2): 1026-1033.

[91] JIA Z Q, LI J J, QU X Y, et al. Enhancing multi-agent systems via reinforcement learning with LLM-based planner and graph-based policy[C]//2025 IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2025: 1240-1246.

[92] WEI Y, SHAN X H, MIAO R, et al. LERO: LLM-driven evolutionary framework with hybrid rewards and enhanced observation for multi-agent reinforcement learning[C]//Advanced Intelligent Computing Technology and Applications. Singapore: Springer, 2025: 15-26.

[93] CHEN R Q, SONG W B, ZU W Q, et al. An LLM-driven framework for multiple-vehicle dispatching and navigation in smart city landscapes[C]//2024 IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2024: 2147-2153.

[94] ZINTGRAF L, SCHULZE S, LU C, et al. VariBAD: Variational bayes-adaptive deep RL via meta-learning[J]. Journal of Machine Learning Research, 2021, 22(289): 1-39.

[95] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proceedings of the 34th International Conference on Machine Learning - Volume 70. New York: ACM, 2017: 1126-1135.

[96] GUPTA A, MENDONCA R, LIU Y X, et al. Meta-reinforcement learning of structured exploration strategies[EB/OL]. (2018-02-20)[2025-10-10]. https://arXiv.org/abs/1802.07245.

[97] RAKELLY K, ZHOU A, QUILLEN D, et al. Efficient off-policy meta-reinforcement learning via probabilistic context variables[EB/OL]. (2019-03-19) [2025-10-10]. https://arXiv.org/abs/1903.08254.

[98] SHARMA N, GHOSH A, MISRA R, et al. Deep meta Q-learning based multi-task offloading in edge-cloud systems[J]. IEEE Transactions on Mobile Computing, 2024, 23(4): 2583-2598.

[99] YUN W J, PARK J, KIM J. Quantum multi-agent meta reinforcement learning[C]//Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence. New York: ACM, 2023: 11087-11095.

[100] MAO W C, QIU H R, WANG C, et al. Multi-agent meta-reinforcement learning: Sharper convergence rates with task similarity[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems, New York: Curran Associates Inc., 2023: 66556-66570.

[101] BOUGZIME O, JABBAR S, CRUZ C, et al. Evaluating Neuro-symbolic AI architectures: Design principles, qualitative benchmark, comparative analysis and results[C]//Conference on Neurosymbolic Learning and Reasoning. Cambridge: PMLR, 2025: 1119-1143.

[102] SHINDO H, DELFOSS Q, DHAMI D S, et al.

BlendRL: A framework for merging symbolic and neural policy learning[EB/OL]. (2025-04-21) [2025-10-10]. https://arXiv.org/abs/2410.11689.

[103] WAN K J, LIU Y T, LIU H Z, et al. A framework for modeling cognitive processes in intelligent agents using behavior trees[C]//Proceedings of the 2025 5th International Conference on Internet of Things and Machine Learning. New York: ACM, 2025: 267-271.

[104] LIU Z C, ZHU Y Y, WANG Z, et al. MIXRTs: Toward interpretable multi-agent reinforcement learning via mixing recurrent soft decision trees[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(5): 4090-4107.

[105] BOGGESS K. Explanations for multi-agent reinforcement learning[C]//Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. New York: ACM, 2025: 29245-29246.

[106] ZHU S C, HAN G J, LIN C, et al. Underwater multiple AUV cooperative target tracking based on minimal reward participation-embedded MARL[J]. IEEE Transactions on Mobile Computing, 2025, 24(5): 4169-4182.

[107] CHEN J M, WANG Y W, WANG J J, et al. Understanding individual agent importance in multi-agent system via counterfactual reasoning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(15): 15785-15794.

[108] RUAN Z H, YU C. Causality-guided exploration for multi-agent reinforcement learning[C]//2024 IEEE International Conference on Agents. Piscataway: IEEE, 2024: 56-59.

[109] MADUMAL P, MILLER T, SONENBERG L, et al. Explainable reinforcement learning through a causal lens[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(3): 2493-2500.

[110] CHU T S, WANG J, CODECÀ L, et al. Multi-agent deep reinforcement learning for large-scale traffic signal control[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(3): 1086-1095.

[111] WANG X Q, KE L J, QIAO Z M, et al. Large-scale traffic signal control using a novel multiagent reinforcement learning[J]. IEEE Transactions on Cybernetics, 2021, 51(1): 174-187.

[112] GU W, KATO S, LIU D B, et al. Integrating suboptimal human knowledge with hierarchical reinforcement learning for large-scale multiagent systems[C]//Advances in Neural Information Processing Systems 37. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 102744-102767.

[113] LI J C, SHI H B, HWANG K S. Using fuzzy logic to learn abstract policies in large-scale multiagent reinforcement learning[J]. IEEE Transactions on Fuzzy Systems, 2022, 30(12): 5211-5224.

[114] LIU Y L, LUO G Y, YUAN Q, et al. GPLight: Grouped multi-agent reinforcement learning for large-scale traffic signal control[C]//Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. California: IJCAI, 2023: 199-207.

[115] HAO Q Y, HUANG W Z, FENG T, et al. GAT-MF: Graph attention mean field for very large scale multi-agent reinforcement learning[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2023: 685-697.

[116] IQBAL S, SHA F. Actor-attention-critic for multi-agent reinforcement learning[EB/OL]. (2019-05-27) [2025-10-10]. https://arXiv.org/abs/1810.02912.

[117] MA C D, LI A M, DU Y L, et al. Efficient and scalable reinforcement learning for large-scale network control[J]. Nature Machine Intelligence, 2024, 6(9): 1006-1020.

[118] ZHAO X Y, WU C. Large-scale machine learning cluster scheduling via multi-agent graph reinforcement learning[J]. IEEE Transactions on Network and Service Management, 2022, 19(4): 4962-4974.

[119] YE Y J, TANG Y, WANG H Y, et al. A scalable privacy-preserving multi-agent deep reinforcement learning approach for large-scale peer-to-peer transactive energy trading[J]. IEEE Transactions on Smart Grid, 2021, 12(6): 5185-5200.

[120] GAO Z, YANG L, DAI Y. Large-scale computation offloading using a multi-agent reinforcement learning in heterogeneous multi-access edge computing[J]. IEEE Transactions on Mobile Computing, 2023, 22(6): 3425-3443.

[121] LEROY P, MORATO P G, PISANE J, et al. IMP-MARL: A suite of environments for large-scale infrastructure management planning via MARL[EB/OL]. (2023-10-27)[2025-10-10]. https://arXiv.org/abs/2306.11551.

[122] RILEY J, CALINESCU R, PATERSON C, et al. Utilising assured multi-agent reinforcement learning within safety-critical scenarios[J]. Procedia Computer Science,

2021, 192: 1061-1070.

[123] GU S D, GRUDZIEN KUBA J, CHEN Y P, et al. Safe multi-agent reinforcement learning for multi-robot control[J]. Artificial Intelligence, 2023, 319: 103905.

[124] HAN S Y, ZHOU S L, WANG J W, et al. A multi-agent reinforcement learning approach for safe and efficient behavior planning of connected autonomous vehicles[EB/OL]. (2022-09-04)[2025-10-10]. https://arXiv.org/abs/2003.04371.

[125] QIU Y B, JIN Y, YU L B, et al. Safe multi-agent reinforcement learning via dynamic shielding[C]//2024 IEEE Conference on Artificial Intelligence. Piscataway: IEEE, 2024: 1254-1257.

[126] LU S T, ZHANG K Q, CHEN T Y, et al. Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(10): 8767-8775.

[127] GU S D, HUANG D Y, WEN M N, et al. Safe multiagent learning with soft constrained policy optimization in real robot control[J]. IEEE Transactions on Industrial Informatics, 2024, 20(9): 10706-10716.

[128] GAO Z, FU J M, JING Z M, et al. MOIPC-MAAC: Communication-assisted multiobjective MARL for trajectory planning and task offloading in multi-UAV-assisted MEC[J]. IEEE Internet of Things Journal, 2024, 11(10): 18483-18502.

[129] CUI J J, LIU Y W, NALLANATHAN A. Multi-agent reinforcement learning-based resource allocation for UAV networks[J]. IEEE Transactions on Wireless Communications, 2020, 19(2): 729-743.

[130] PAN Y H, WANG X C, XU Z Y, et al. GNN-empowered effective partial observation MARL method for AoI management in multi-UAV network[J]. IEEE Internet of Things Journal, 2024, 11(21): 34541-34553.

[131] SHI R Y, YU X, WANG Y D, et al. Symmetry-informed MARL: A decentralized and cooperative UAV swarm control approach for communication coverage[J]. IEEE Transactions on Mobile Computing, 2025, 24(9): 8039-8056.

[132] ZHANG Y, MOU Z Y, GAO F F, et al. UAV-enabled secure communications by multi-agent deep reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2020, 69(10): 11599-11611.

[133] CHEN S T, LIU G J, ZHOU Z Y, et al. Robust multi-agent reinforcement learning method based on adversarial domain randomization for real-world dual-UAV cooperation[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(1): 1615-1627.

[134] XIA Z Y, DU J, WANG J J, et al. Multi-agent reinforcement learning aided intelligent UAV swarm for target tracking[J]. IEEE Transactions on Vehicular Technology, 2022, 71(1): 931-945.

[135] CHEN D Z, QI Q, FU Q L, et al. Transformer-based reinforcement learning for scalable multi-UAV area coverage[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(8): 10062-10077.

[136] BIAGIONI D, ZHANG X Y, WALD D, et al. PowerGridworld: A framework for multi-agent reinforcement learning in power systems[C]//Proceedings of the Thirteenth ACM International Conference on Future Energy Systems. New York: ACM, 2022: 565-570.

[137] WANG J H, XU W K, GU Y J, et al. Multi-agent reinforcement learning for active voltage control on power distribution networks[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. New York: ACM, 2021: 3271-3284.

[138] CHEN D, CHEN K A, LI Z J, et al. PowerNet: Multi-agent deep reinforcement learning for scalable powergrid control[J]. IEEE Transactions on Power Systems, 2022, 37(2): 1007-1017.

[139] SHARMA M K, ZAPPONE A, ASSAAD M, et al. Distributed power control for large energy harvesting networks: A multi-agent deep reinforcement learning approach[J]. IEEE Transactions on Cognitive Communications and Networking, 2019, 5(4): 1140-1154.

[140] ROESCH M, LINDER C, ZIMMERMANN R, et al. Smart grid for industry using multi-agent reinforcement learning[J]. Applied Sciences, 2020, 10(19): 10196900.

[141] YU T, WANG H Z, ZHOU B, et al. Multi-agent correlated equilibrium Q(λ) learning for coordinated smart generation control of interconnected power grids[J]. IEEE Transactions on Power Systems, 2015, 30(4): 1669-1679.

[142] MU C X, LIU Z Y, YAN J, et al. Graph multi-agent reinforcement learning for inverter-based active voltage control[J]. IEEE Transactions on Smart Grid, 2024, 15(2): 1399-1409.

[143] HU D E, YE Z H, GAO Y Q, et al. Multi-agent deep reinforcement learning for voltage control with coordinated active and reactive power optimization[J]. IEEE Transactions on Smart Grid, 2022, 13(6): 4873-4886.

[144] GAO Y Q, WANG W, YU N P. Consensus multi-agent reinforcement learning for volt-VAR control in power

distribution networks[J]. IEEE Transactions on Smart Grid, 2021, 12(4): 3594-3604.
[145] LIU T Y, CHEN H C, HU J F, et al. Generalized multi-agent competitive reinforcement learning with differential augmentation[J]. Expert Systems with Applications, 2024, 238: 121760.
[146] DASKALAKIS C, FOSTER D J, GOLOWICH N. Independent policy gradient methods for competitive reinforcement learning[EB/OL]. (2021-01-11)[2025-10-10]. https://arXiv.org/abs/2101.04233.
[147] CHEN C Q, YANG H N, ZHAI C J, et al. Competitive pricing for ride-sourcing platforms with MARL[J]. Transportation Research Part C: Emerging Technologies, 2024, 165: 104697.
[148] WU J H, WANG J D, KONG X Y. Strategic bidding in a competitive electricity market: An intelligent method using Multi-Agent Transfer Learning based on reinforcement learning[J]. Energy, 2022, 256: 124657.
[149] LIU Z, LU M, WANG Z, et al. Welfare maximization in competitive equilibrium: Reinforcement learning for markov exchange economy[C]//International Conference on Machine Learning. Cambridge: PMLR, 2022: 13870-13911.
[150] BAI Y, JIN C. Provable self-play algorithms for competitive reinforcement learning[EB/OL]. (2020-07-09)[2025-10-10]. https://arXiv.org/abs/2002.04017.

作者简介

韩光洁 男,1972年8月出生于黑龙江省绥化市. 现为河海大学信息科学与工程学院教授、博士生导师. 主要研究方向为水声通信与组网、水利智能物联网、人工智能、网络与安全等. 中国电子学会会员编号:E190157962M.
E-mail: hanguangjie@gmail.com

朱胜超 男,2001年9月出生于山东省德州市. 现为河海大学计算机与软件学院博士研究生. 主要研究方向为多智能体强化学习、软件定义网络、智慧海洋. 中国电子学会会员编号:E190197863A.
E-mail: zhushengchao77@gmail.com

林 川 男,1988年2月出生于辽宁省丹东市. 现为东北大学软件学院副教授、博士生导师. 主要研究方向为多智能体强化学习、软件定义网络、智慧海洋等.
E-mail: chuanlin1988@gmail.com

江金芳 女,1988年1月出生于安徽省六安市. 现为河海大学信息科学与工程学院教授、博士生导师. 主要研究方向为水下通信与组网、水下信任等. 中国电子学会会员编号:E190157961M.
E-mail: jiangjinfang@hhu.edu