

## 基于多智能体强化学习的博弈综述

李艺春<sup>1,2</sup> 刘泽娇<sup>1,3</sup> 洪艺天<sup>1,2</sup> 王继超<sup>1,2</sup> 王健瑞<sup>1,2</sup> 李毅<sup>1,2</sup> 唐漾<sup>1,2</sup>

**摘要** 多智能体强化学习 (Multi-agent reinforcement learning, MARL) 作为博弈论、控制论和多智能体学习的交叉研究领域, 是多智能体系统 (Multi-agent systems, MASs) 研究中的前沿方向, 赋予智能体在动态多维的复杂环境中通过交互和决策完成多样化任务的能力. 多智能体强化学习正在向应用对象开放化、应用问题具身化、应用场景复杂化的方向发展, 并逐渐成为解决现实世界中博弈决策问题的最有效工具. 本文对基于多智能体强化学习的博弈进行系统性综述. 首先, 介绍多智能体强化学习的基本理论, 梳理多智能体强化学习算法与基线测试环境的发展进程. 其次, 针对合作、对抗以及混合三种多智能体强化学习任务, 从提高智能体合作效率、提升智能体对抗能力的维度来介绍多智能体强化学习的最新进展, 并结合实际应用探讨混合博弈的前沿研究方向. 最后, 对多智能体强化学习的应用前景和发展趋势进行总结与展望.

**关键词** 多智能体强化学习, 多智能体系统, 博弈决策, 均衡求解

**引用格式** 李艺春, 刘泽娇, 洪艺天, 王继超, 王健瑞, 李毅, 唐漾. 基于多智能体强化学习的博弈综述. 自动化学报, 2025, 51(3): 540-558

**DOI** 10.16383/j.aas.c240478 **CSTR** 32138.14.j.aas.c240478

### Multi-agent Reinforcement Learning Based Game: A Survey

LI Yi-Chun<sup>1,2</sup> LIU Ze-Jiao<sup>1,3</sup> HONG Yi-Tian<sup>1,2</sup> WANG Ji-Chao<sup>1,2</sup>  
WANG Jian-Rui<sup>1,2</sup> LI Yi<sup>1,2</sup> TANG Yang<sup>1,2</sup>

**Abstract** Multi-agent reinforcement learning (MARL), which stands at the intersection of game theory, cybernetics and multi-agent learning, represents the cutting-edge domain within the realm of multi-agent systems (MASs) research. MARL empowers the agents with the capability to complete a variety of complex tasks through interaction and decision-making in the dynamic multi-dimensional and complicated practical environment. When progressing towards the openness of application objects, the embodiment of application issues and the complication of application contexts, MARL is gradually becoming the most effective tool for solving game and decision-making problems in the real world. This paper systematically reviews the game based on MARL. First, the basic theory of MARL is introduced, and the development process of MARL algorithms and the baseline testing environment have been introduced and summarized. Then, we focus on three types of tasks within MARL, which are cooperation, competition and mixed tasks. The latest progress in MARL is introduced by concentrating on improving the cooperative efficiency and enhancing the adversarial abilities of agents, and the most recent researches on mixed games, in combination with their practical applications, are investigated. Finally, the prospects of application and the trends of development for MARL are summarized and prospected.

**Key words** Multi-agent reinforcement learning (MARL), multi-agent systems (MASs), game and decision-making, equilibrium solution

**Citation** Li Yi-Chun, Liu Ze-Jiao, Hong Yi-Tian, Wang Ji-Chao, Wang Jian-Rui, Li Yi, Tang Yang. Multi-agent reinforcement learning based game: A survey. *Acta Automatica Sinica*, 2025, 51(3): 540-558

收稿日期 2024-07-05 录用日期 2024-10-16

Manuscript received July 5, 2024; accepted October 16, 2024

国家自然科学基金 (62233005, U2441245), 中国博士后科学基金 (2024M750904) 资助

Supported by National Natural Science Foundation of China (62233005, U2441245) and China Postdoctoral Science Foundation (2024M750904)

本文责任编辑 温广辉

Recommended by Associate Editor WEN Guang-Hui

1. 华东理工大学能源化工过程智能制造教育部重点实验室 上海 200237 2. 华东理工大学信息科学与工程学院 上海 200237 3. 华东理工大学数学学院 上海 200237

1. Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237 2. School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237 3. School of Mathematics, East China University of Science and Technology, Shanghai 200237

人工智能作为新一轮科技革命和产业变革的重要驱动力量, 是经济发展的新引擎、社会发展的加速器. 以 SORA、ChatGPT 为代表的大模型革新和崛起正引领着新一轮全球人工智能技术的发展浪潮, 并推动人工智能技术向通用人工智能迈进<sup>[1]</sup>. 多智能体强化学习 (Multi-agent reinforcement learning, MARL) 作为人工智能领域中的一种关键行为决策和控制技术, 是人工智能驱动自动化技术发展的重要助推力, 也是研究群体智能系统的关键工具. 多智能体系统 (Multi-agent systems, MASs) 作为群体智能的主要研究对象, 是由多个具有博弈行为的智能体组成的复杂系统, 其中每个智能体在交互

决策时具有一定的自主、独立和协作学习能力,也是自主智能系统在国防安全、能源系统、工业制造、智能交通等领域的典型代表<sup>[2-6]</sup>。有关多智能体系统的研究呈现出交叉融合的趋势,是新一代人工智能的前沿热点方向<sup>[7-8]</sup>。

多智能体强化学习作为博弈论、控制论和多智能体学习的交叉研究领域,将强化学习和博弈论的思想应用于多智能体系统的学习、优化和控制中,为解决多智能体系统的复杂决策问题提供有效的方法与技术手段<sup>[7, 9-10]</sup>。其中,强化学习是一种经典的机器学习方法,智能体通过与环境不断交互来实现长期累积回报最大化<sup>[2, 4, 11]</sup>。强化学习具有无需预先精准建模和处理高维非线性数据等优点,已在解决各类动态决策问题中取得巨大成功<sup>[12-13]</sup>,但将强化学习应用于多智能体博弈场景时,智能体决策会受到其他智能体策略变化的影响,进而导致强化学习的环境稳定性条件被打破<sup>[14-15]</sup>。动态环境的部分可观性、不确定性和随机性增加了智能体博弈决策的难度<sup>[16-20]</sup>,复杂博弈场景也为环境建模、奖励设计和策略表达带来额外的挑战<sup>[21-22]</sup>。多智能体强化学习的应用对象开放化、应用问题具身化、应用场景复杂化,对模型和算法的可扩展性提出了更高的要求。应用对象开放化是指多智能体强化学习面向开源、异构的应用对象,随着开放环境中新样本、新态势的出现,智能体数量动态增加或减少,多智能体强化学习的应用对象可适应不确定与动态变化的开放环境的要求。应用问题具身化强调依赖智能体的物理形态,并非是抽象的算法运算和数据处理问题,可实现智能博弈从基于算法的虚拟决策到依赖现实物理形态的实际控制的突破,而传统的博弈智能方法尚未真正解决智能体与真实物理世界交互的难题<sup>[3, 23-24]</sup>。应用场景复杂化意味着多智能体强化学习的应用场景从单一的棋牌类博弈<sup>[25]</sup>、战略游戏<sup>[26]</sup>转向大规模智能无人集群、智能交通管控、物流调度管理、广告竞价拍卖等复杂应用中<sup>[15]</sup>,实际应用场景愈加复杂化、多样化。

在多智能体强化学习中智能体以任务为导向,根据任务不同可将其划分为合作、对抗和混合任务,这分别与博弈论中合作、对抗和混合博弈相对应。因此,本文聚焦于上述三种博弈类型进行讨论。以人机协同问题为例<sup>[2]</sup>,合作博弈致力于探索如何通过智能体间的合作来实现共同目标<sup>[27-28]</sup>。以棋牌类对抗游戏为典型代表<sup>[17-18]</sup>的对抗博弈则聚焦于竞争环境中的策略制定,以获得竞争优势或确保个体利益的最大化<sup>[29-31]</sup>。混合博弈则融合了合作与对抗的元素,是包含智能体合作与对抗关系的博弈模型,更加贴合现实世界的复杂交互场景<sup>[32-34]</sup>,例如公共

能源分配<sup>[4]</sup>等问题。随着多智能体强化学习计算性能和算法的不断优化,多智能体博弈决策技术也获得了快速发展。随着以强化学习为核心算法的 AlphaGo 在与人类对弈的围棋比赛中取得惊人成绩<sup>[35]</sup>,多智能体博弈决策技术受到研究者的广泛关注,有关应用领域从围棋、德州扑克<sup>[25]</sup>等棋牌类的分步对抗游戏拓展到雷神之锤<sup>[36]</sup>、星际争霸<sup>[37-38]</sup>和王者荣耀<sup>[26]</sup>等多人实时战略游戏,以及面向实际场景的多机协同作战<sup>[2]</sup>、自动驾驶<sup>[5]</sup>和多机器人协作<sup>[39]</sup>。

多智能体博弈场景的不断扩展对基于多智能体强化学习的博弈求解方法提出了新的要求,同时多智能体强化学习的新理论、新算法的不断涌现极大地促进了多智能体博弈决策迈向更广阔的研究空间。因此,本文聚焦于多智能体强化学习下博弈的最新进展,探讨多智能体强化学习的前沿问题与发展动态。此外,关于多智能体系统中博弈模型分类和经典求解方法<sup>[7, 15]</sup>、强化学习的基础理论、典型算法<sup>[40-41]</sup>,以及针对某类博弈类型或学习问题的研究分析<sup>[17-18, 42-44]</sup>等方面,已有许多优秀、经典的文献综述,为广大研究者提供了启发式的思想指导。本文涉及的相关理论知识等内容可参阅上述文献中更细致的介绍。

本文在现有综述和最新研究成果的基础上,聚焦于合作、对抗以及混合的三种多智能体任务,重点梳理相应的博弈模型、智能体交互决策的核心问题以及基于多智能体强化学习的最新求解方法。第 1 节介绍单智能体强化学习、多智能体强化学习及其对应的数学模型和算法、学习方式,概述安全、鲁棒等多智能体强化学习的不同分支,梳理最新的多智能体强化学习的测试环境。区别于现有综述,第 2 节针对合作、对抗博弈分别从提高智能体合作效率、提升智能体对抗能力的维度来介绍多智能体强化学习的最新进展,并结合现实世界中的实际应用探讨混合博弈的前沿研究方向。第 3 节总结多智能体博弈决策的现阶段挑战。最后,第 4 节结合人工智能发展的前沿热点,展望多智能体系统博弈决策中理论与实际应用的未来发展。

## 1 多智能体强化学习理论

得益于人工智能的兴起和深度学习中的技术突破,研究人员从算法收敛性和可扩展性、智能体学习效率、博弈策略求解和实际应用场景等方面对多智能体强化学习算法和基线测试环境进行完善和创新。多智能体强化学习算法从低维离散空间向高维连续空间扩展,研究规模从少量智能体参与向大规模集群场景不断发展。图 1 梳理了多智能体强化学习算法与测试环境的发展历程。

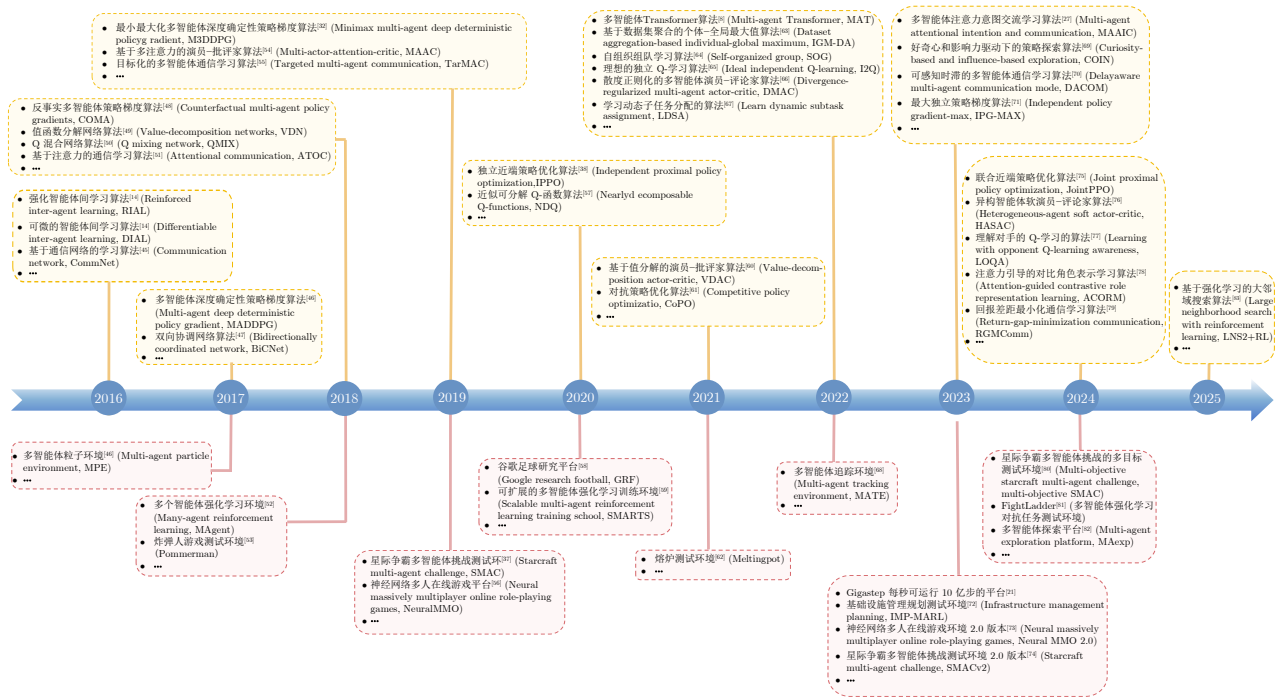


图 1 多智能体强化学习算法与基线测试环境的发展进程

Fig. 1 Development process of algorithms and baseline test environments of multi-agent reinforcement learning

## 1.1 单智能体强化学习

强化学习的前期工作主要聚焦于单个智能体的序贯决策问题,即单智能体强化学习.单智能体强化学习常被建模为马尔科夫决策过程<sup>[11]</sup> (Markov decision process, MDP).标准的马尔科夫决策过程用五元组  $\langle S, A, P, R, \gamma \rangle$  表示,五个变量分别代表环境状态集合、动作集合、状态转移函数、奖励函数和折扣因子,其中,  $\gamma \in [0, 1]$  用于描述智能体在决策过程中对即时奖励与未来奖励的关注度.单智能体序贯决策的过程可表述为:在时间步  $t$ ,假定环境处于状态  $s_t \in S$ ,智能体根据其策略  $\pi$  执行动作  $a_t \in A$ ,该动作使得环境根据转移概率函数转移到下一个状态  $s_{t+1} \sim P(\cdot | s_t, a_t)$ ,并根据奖励函数向智能体返回一个即时的奖励  $r(s_t, a_t, s_{t+1})$ .该过程随着时间重复进行,状态转移满足马尔科夫性,进而形成马尔科夫决策过程<sup>[7]</sup>.

单智能体强化学习的目标是使得智能体在任意状态  $s \in S$  下找到策略  $\pi$  来最大化状态值函数  $V_\pi(s)$ ,即最大化如下期望累积折扣奖励<sup>[9]</sup>

$$V_\pi(s) = \mathbb{E}_{a_t \sim \pi(s_t), s_{t+1} \sim P(\cdot | s_t, a_t)} \left( \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \mid s_0 = s \right), \forall s \in S \quad (1)$$

求解马尔科夫决策过程的传统方法是基于动态规划的思想进行值迭代或策略迭代<sup>[84]</sup>,例如通过值迭代获得最优策略  $\pi^*$  对应的状态值函数  $V_{\pi^*}(s)$ .  $V_{\pi^*}(s)$  满足如下 Bellman 最优方程,

$$V_{\pi^*}(s) = \max_{a \in A} \mathbb{E}_{s' \sim P(\cdot | s, a)} (r(s, a, s') + \gamma V_{\pi^*}(s') \mid s, a) = \max_{a \in A} Q_{\pi^*}(s, a), \forall s \in S \quad (2)$$

其中,  $Q_{\pi^*}(s, a)$  为最优策略  $\pi^*$  对应的状态-动作值函数,它表示在状态  $s$  下采取某个动作  $a$  后所获得的期望回报.

基于动态规划等传统方法来求解马尔科夫决策过程的主要局限在于在高维空间中计算复杂度的急剧增加,并且只适用于状态转移概率函数和奖励函数等环境模型已知的情况.然而在多机协同作战<sup>[2]</sup>、自动驾驶<sup>[5]</sup>等大多数实际场景中人们无法得知准确的模型信息,因此,需要利用基于采样与环境进行交互学习的无模型强化学习方法求解马尔科夫决策过程<sup>[7]</sup>.无模型强化学习方法主要分为:基于值函数<sup>[84]</sup>、基于策略<sup>[85]</sup>以及演员-评论家 (Actor-critic, AC) 算法<sup>[86]</sup>.基于值函数的方法以 Q-learning 算法为代表,它基于时序差分 (Temporal-difference, TD) 的思想,迭代并更新对 Q 值的估计,从而推导出最优策略<sup>[84]</sup>.基于策略的方法通过对策略进行参数化来

直接对策略空间进行搜索, 从而找到最优策略, 相比于值函数的方法更适用于处理复杂连续动作空间. AC 算法同时估计值函数和策略函数, 是基于值函数与基于策略的方法的有机融合, 已成为主流的强化学习算法, 其本质是基于策略的方法, 但可以解决基于策略的方法带来的策略梯度方差较大、采样效率低下等问题<sup>[41]</sup>. 经典的无模型强化学习方法的详细介绍可参考文献 [7, 15] 等.

## 1.2 多智能体强化学习

马尔科夫博弈 (Markov game, MG) 是描述多个智能体序贯决策过程的常用数学模型<sup>[9, 87]</sup>. 马尔科夫博弈是马尔科夫决策过程在多智能体系统中的扩展, 也常被称为随机博弈 (Stochastic game, SG)<sup>[88-89]</sup>. 马尔科夫博弈一般由六元组  $\langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}, \mathcal{P}, \{r_i\}, \gamma \rangle$  表示<sup>[90]</sup>,  $i \in \mathcal{N} = \{1, \dots, N\}$ ,  $N$  代表智能体个数,  $\mathcal{S}$  为所有智能体所处环境的状态集合,  $\mathcal{A}_i$  为智能体  $i$  的可执行动作集合, 记联合动作空间与转移概率函数分别表示为  $\mathbf{A} := \prod_{i \in \mathcal{N}} \mathcal{A}_i$ ,  $\mathcal{P} : \mathcal{S} \times \mathbf{A} \times \mathcal{S} \rightarrow \Delta(\mathcal{S})$ ,  $\mathcal{P}$  衡量的是在某个时间步  $t$ , 所有智能体执行联合动作  $\mathbf{a}_t \in \mathbf{A}$  后, 环境从当前状态  $s_t \in \mathcal{S}$  转移到新的状态  $s_{t+1} \in \mathcal{S}$  的概率,  $r_i : \mathcal{S} \times \mathbf{A} \times \mathcal{S} \rightarrow \mathbf{R}$  为智能体  $i$  的即时奖励函数. 多智能体强化学习的目的是让每个智能体  $i$  学习一个策略  $\pi_i$ , 该策略可以最大化如下期望累积折扣奖励<sup>[7]</sup>

$$V_{\pi_i, \pi_{-i}}(s) = \mathbf{E}_{\mathbf{a}_t \sim (\pi_i, \pi_{-i})(s_t), s_{t+1} \sim \mathcal{P}(\cdot | s_t, \mathbf{a}_t)} \left( \sum_{t=0}^{\infty} \gamma^t \times r_i(s_t, \mathbf{a}_t, s_{t+1}) \mid s_0 = s \right) \quad (3)$$

其中, 下标  $-i$  表示除智能体  $i$  之外的其他智能体集合, 即  $\mathcal{N} \setminus \{i\}$ , 对应地,  $\pi_{-i}$  代表除智能体  $i$  之外的智能体策略组合. 在多智能体系统中, 智能体会受到环境和其他智能体的影响, 并且环境状态的转移以及每个智能体收到的奖励都由所有智能体的联合动作决定. 因此, 单智能体强化学习与多智能体强化学习的策略求解存在本质差异<sup>[48, 54]</sup>.

现有文献从不同角度对多智能体强化学习范式进行分类, 例如文献 [40] 从学习方法角度将其分为联合学习和独立学习, 从通信角度将其分为集中式学习、分散式学习和分布式学习. 本文从算法训练与动作执行模式的角度出发, 将多智能体强化学习范式分为如下三类: 分布式训练分散式执行<sup>[91]</sup> (Distributed training with decentralized execution, DTDE)、集中式训练集中式执行<sup>[75]</sup> (Centralized training with centralized execution, CTCE) 和集

中式训练分散式执行<sup>[49-50, 92]</sup> (Centralized training and decentralized execution, CTDE). 在 DTDE 学习范式中, 每个智能体学习自己的策略, 拥有独立的模型, 独立地与环境进行交互. 在 CTCE 学习范式中, 执行时所有的个体都完全由唯一的中心控制器控制, 在训练时将所有智能体的观测和动作空间分别联合起来, 建模成一个联合的智能体进行训练. 而在 CTDE 学习范式中, 在训练时集中式地对所有智能体的联合策略进行评估, 显式地考虑其他所有智能体的局部观测以及动作; 在执行时每个智能体分散单独执行, 每个智能体依靠自身局部观测和通信信息进行动作选择<sup>[63]</sup>. 有关多智能体强化学习中不同学习范式的具体介绍和相关算法可参考文献 [15, 40, 42] 等.

此外, 智能体交互的多样性、复杂性和与现实任务贴合的紧密性, 为多智能体强化学习的算法设计和实际应用带来挑战. 相关研究聚焦于解决如下问题: 从多智能体系统角度, 受限于智能体的局部感知力与物理空间中的距离而导致的部分可观性<sup>[7, 91]</sup>、各个智能体的策略不断变化使得整个系统的环境对于每个智能体来说呈现出非平稳的特性<sup>[14, 34]</sup>、智能体之间的差异体现出异构或异质性<sup>[76, 93]</sup> 以及存在动作空间和规模可扩展的智能体, 上述情况使得联合动作空间呈指数级增长, 进而带来维度灾难性等问题<sup>[17, 31]</sup>. 从强化学习算法角度, 除了存在探索与利用的平衡<sup>[69, 94-97]</sup>、智能体奖励稀疏性<sup>[90, 98]</sup>、信用分配不合理<sup>[99-100]</sup> 等典型问题外, 通用人工智能的发展也对强化学习中决策的可解释性<sup>[7, 44, 101]</sup>、策略的鲁棒性<sup>[32, 102]</sup> 以及算法的通用性<sup>[76]</sup> 提出了更高的要求.

## 1.3 其他多智能体强化学习分支

随着多学科深度交叉融合, 强化学习、博弈论与控制论领域中涌现出如安全、鲁棒以及元强化学习等方法, 进而被应用于解决各种现实世界中的复杂问题<sup>[103]</sup>. 由此, 多智能体强化学习领域也分化出多个不同的分支<sup>[7]</sup>.

安全多智能体强化学习是基于安全强化学习的多智能体学习方法<sup>[43]</sup>. 安全强化学习一直是人工智能领域中的热点方向<sup>[104]</sup>, 传统的强化学习方法只关注期望奖励最大化, 没有考虑满足安全约束条件, 一味追求奖励最大化则会给机器人协作、智能交通和工业过程控制等应用带来巨大隐患. 安全强化学习是指满足一定安全约束的最大化长期回报的学习过程<sup>[105]</sup>, 常被定义为一个具有状态约束的马尔科夫决策过程. 在多智能体系统中, 智能体需要考虑自身和其他智能体的安全约束<sup>[104]</sup>, 避免因智能体的随

机探索而违反安全规则,这增加了整个系统优化过程的复杂性,为开发遵循约束的安全多智能体强化学习算法带来挑战<sup>[43, 106]</sup>.

鲁棒多智能体强化学习将鲁棒强化学习扩展到多智能体系统中,它常用于解决多智能体场景中受到扰动或存在不确定性的序贯决策问题<sup>[7]</sup>.鲁棒强化学习在标准的强化学习算法中融合了鲁棒优化的思想<sup>[7]</sup>,旨在提升强化学习算法在面对干扰或不确定性时的鲁棒性<sup>[102]</sup>.在对抗博弈中,鲁棒多智能体强化学习不仅注重于考虑对方智能体策略的变化对我方决策的影响<sup>[32]</sup>,也希望在面向对抗性攻击、噪声或恶意策略扰动时保证算法的鲁棒性或者学习强鲁棒性的策略<sup>[107]</sup>.

元多智能体强化学习是基于元学习的多智能体强化学习方法,它将元学习中的单任务框架扩展到多任务框架,利用元学习中使用现有任务的数据来学习或表示模型参数的思想,可以使用更少的样本适应更广泛的任务并迁移已有知识<sup>[77, 108]</sup>.元多智能体强化学习在解决零样本 (Zero-shot) 和小样本 (Few-shot) 问题中已取得显著成效<sup>[109]</sup>.元多智能体强化学习现有工作主要围绕多任务场景下算法的泛化能力展开研究,探讨如何适应未见过的任务和场景,从而构建具有较好收敛性保障与泛化性能的强化学习算法<sup>[108]</sup>.

此外,提高决策可解释性的因果多智能体强化学习<sup>[44]</sup>、注重数据隐私保护的联邦多智能体强化学习<sup>[110]</sup>与根据专家样本求解最优策略的逆多智能体强化学习<sup>[111]</sup>等算法也引起研究人员的广泛关注,为解决多智能体博弈决策问题提供了新的思路.

#### 1.4 多智能体强化学习基线测试环境

多智能体强化学习的基线测试环境是研究多智能体博弈的基石,为多智能体强化学习算法验证、模拟智能体博弈效果提供公平标准的测试平台一直

是研究人员的关注重点<sup>[52-53, 56, 62]</sup>.主流的基线测试环境包括:适用于合作导航、通信交流和追捕猎物等多种任务的多智能体粒子环境<sup>[46]</sup> (Multi-agent particle environment, MPE)、采用游戏设计理念的合作星际争霸多智能体挑战测试环境<sup>[37-38]</sup> (Starcraft multi-agent challenge, SMAC)、基于物理的 3D 足球模拟器的谷歌足球研究平台<sup>[58]</sup> (Google research football, GRF)、模拟现实世界中自动驾驶的可扩展多智能体强化学习训练环境<sup>[59]</sup> (Scalable multi-agent reinforcement learning training school, SMARTS) 等.

近期,研究者们针对可扩展性、任务多样性和计算性能等方面对原有测试环境进行了改进. Neural MMO 2.0<sup>[73]</sup> 允许用户定义多种目标和奖励,用于训练在面对未知任务、地图和对手时仍能表现出良好泛化能力的智能体; Gigastep<sup>[21]</sup> 能够在消费级硬件上每秒执行十亿个环境步数,极大提升了计算性能,并适用于对大规模智能体 (大于 1000) 进行测试; SMA-Cv2<sup>[74]</sup> 引入随机性和动态变化的任务环境,突破了原有 SMAC 环境中确定性、完全可观的限制.随着多种测试环境的不断开发,高质量的数据集也随之出现.例如,同时涵盖离线强化学习和多智能体离线强化学习的数据集 Hokoff<sup>[22]</sup>,它包含了不同难度级别的任务,用于模拟与现实世界中不同水平的对手进行对抗.表 1 总结了最新的多智能体强化学习测试环境.随着多智能体博弈决策技术的发展,研究者们期待未来可以在满足强大算力需求下设计轻量级、标准化的开源基线环境来验证多种算法的鲁棒性、可扩展性,并缩小与实际博弈场景的差距.

## 2 最新进展

根据智能体之间的博弈关系,可将多智能体系统中的博弈分为合作、对抗以及包含合作与对抗的混合博弈,这三种博弈类型分别对应多智能体强化

表 1 多智能体强化学习的最新测试环境介绍

Table 1 Introduction of the latest test environments of multi-agent reinforcement learning

测试环境	任务类型	适用场景/特点	动作空间	
			连续	离散
MATE <sup>[68]</sup> (2022)	混合	针对多智能体目标覆盖控制,如无线传感器网络	✓	✓
Gigastep <sup>[21]</sup> (2023)	混合	支持具有随机性和部分可观性的 3D 动态环境	✓	✓
IMP-MARL <sup>[72]</sup> (2023)	合作	针对基础设施管理规划,如海上风力发电机组维护		✓
Neural MMO 2.0 <sup>[73]</sup> (2023)	混合	在 Neural MMO 环境中增加自定义的目标和奖励		✓
SMACv2 <sup>[74]</sup> (2023)	合作	在 SMAC 环境中增加随机性和部分可观察性		✓
Multi-objective SMAC <sup>[80]</sup> (2024)	混合	在 SMAC 环境中增加长期任务和多个对抗目标		✓
FightLadder <sup>[81]</sup> (2024)	对抗	针对多种跨平台视频格斗游戏,如街霸、拳皇		✓
MAexp <sup>[82]</sup> (2024)	合作	用于多规模、多类型机器人团队合作探索策略	✓	

学习中的合作、对抗以及混合任务. 马尔科夫博弈是具有合作或竞争目标的多智能体交互决策的博弈论框架. 多智能体强化学习是一种高效且通用的马尔科夫博弈求解方法, 可解决多个智能体在共享随机环境中的序贯决策问题. 本节将基于马尔科夫博弈对多智能体环境下合作、对抗和混合博弈中的最新工作进行介绍.

## 2.1 合作博弈

在合作博弈中, 所有智能体共享相同的全局奖励函数, 需要相互合作来最大化全局累计奖励. 合作多智能体强化学习旨在为多个智能体训练对应的合作策略, 例如多机器人协作<sup>[39]</sup>和无人机导航<sup>[106]</sup>等应用场景, 从而使得智能体能够相互合作以完成共同的目标任务<sup>[64-65]</sup>.

### 2.1.1 合作博弈模型与学习方法

由于局部可观性是现实世界中的普遍约束, 因此, 多智能体强化学习合作任务中的马尔科夫博弈常用去中心化部分可观的马尔科夫决策过程<sup>[66, 78, 112]</sup> (Decentralized partially observable MDP, Dec-POMDP) 这一数学模型来描述, 它可以定义为一个元组  $\langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}, \mathcal{P}, R, \Omega, \gamma \rangle$ ,  $i \in \mathcal{N}$ , 其中, 观测函数  $\Omega(s, i)$  代表每个智能体从状态到观测的映射, 其余符号含义与马尔科夫博弈中的定义相同. 去中心化部分可观的马尔科夫决策过程中, 每个智能体  $i$  的目标是找到最优联合策略  $\pi^*$  以最大化总期望收益  $J(\pi) = E(\sum_{t=0}^{\infty} \gamma^t r(s_t, \mathbf{a}_t))$ <sup>[66]</sup>. 此外, 研究者也常利用网络化多智能体马尔科夫决策过程 (Networked multi-agents MDP, N-MMDP) 对多智能体合作任务进行建模<sup>[113]</sup>.

在多智能体合作博弈中, 文献 [15] 将多智能体强化学习范式分为独立学习、联合学习、基于协作图和 CTDE 等方式. 独立学习直接将单智能体方法独立地应用到每一个智能体, 容易造成学习过程中的非平稳性, 适用于问题设置简单的多智能体合作任务<sup>[91]</sup>. 联合学习将整个多智能体系统视为一个单智能体, 将单智能体强化学习方法迁移到这个单智能体中, 然而这会加剧智能体策略探索与利用的不平衡<sup>[94]</sup>. 基于协作图的方法利用图来刻画智能体之间的交互关系, 同时分解联合  $Q$  值函数, 虽然它可以结合独立学习和联合学习的优点, 但在实际应用中难以获得准确的协作图<sup>[114]</sup>. 由于 CTDE 可以在一定程度上解决多智能体强化学习中非平稳性<sup>[92]</sup>与信用分配问题<sup>[60, 100]</sup>, 也能同时学习协作图中对联合  $Q$  值函数的分解<sup>[15]</sup>, 本节将主要围绕 CTDE 框架下的合作多智能体强化学习方法进行介绍.

遵循 CTDE 框架的典型多智能体强化学习算法可分为如下两类: 基于值函数分解<sup>[27, 49-50, 57, 67]</sup>与基于 AC 的算法<sup>[66, 76, 92]</sup>. 在基于值函数分解的算法中, VDN<sup>[49]</sup> (Value-decomposition networks) 和 QMIX<sup>[50]</sup> 是最典型的算法. VDN 在值函数可分解的假设下要求联合值函数是每个智能体值函数的线性和, 即

$$Q^{\text{tot}}(\boldsymbol{\tau}, \mathbf{a}) = \sum_{i=1}^N Q_i(\tau_i, a_i) \quad (4)$$

其中,  $\tau_i$  代表智能体  $i$  自身的局部动作-观测历史,  $\boldsymbol{\tau}$  代表联合动作-观测历史. 由于 VDN 算法在部分场景下因线性假设而受限, 使用神经网络来近似联合值函数的算法 QMIX 由此被提出. QMIX 在  $Q^{\text{tot}}(\boldsymbol{\tau}, \mathbf{a})$  与每个  $Q_i(\tau_i, a_i)$  之间施加了如下单调性约束

$$\frac{\partial Q^{\text{tot}}(\boldsymbol{\tau}, \mathbf{a})}{\partial Q_i(\tau_i, a_i)} \geq 0 \quad (5)$$

此外, 不同于 VDN 与 QMIX 算法分解全局动作值, 值分解演员-评论家算法 (Value-decomposition actor-critic, VDAC) 将值分解的思想与基于 AC 的方法相结合对全局状态值进行分解<sup>[60]</sup>, 并引入局部状态值与全局状态值之间的单调关系, 进而改善了信用分配问题.

在 CTDE 范式中, 全局  $Q$  值是由所有智能体的集中训练得到的<sup>[46]</sup>, 从而导致智能体的行为容易受到其他智能体次优行为的影响. 因此, 应用 CTDE 范式需要满足个体-全局最大值<sup>[115]</sup> (Individual-global maximum, IGM) 原则, 即

$$\arg \max_{\mathbf{a}} Q(\boldsymbol{\tau}, \mathbf{a}) = \begin{bmatrix} \arg \max_{a_1} q_1(\tau_1, a_1) \\ \vdots \\ \arg \max_{a_N} q_N(\tau_N, a_N) \end{bmatrix} \quad (6)$$

从而使得集中式策略与分散式策略一致<sup>[40]</sup>. 然而在满足 IGM 原则的算法中, 常面临有损分解问题, 即从全局  $Q$  值到单个  $q_i$  值的分解总存在误差, 产生的误差会在值迭代过程中不断累积, 导致智能体学习效率下降. 文献 [63] 提出的 DAgger-based IGM 算法将 IGM 中的有损分解从迭代过程中分离出来以避免误差累积, 提升了算法在智能体合作任务中的性能.

### 2.1.2 提高智能体合作效率

虽然 CTDE 范式是合作多智能体强化学习中的主流学习范式, 但该范式仍存在一些问题, 例如, CTDE 中存在一个中心化评估器<sup>[46]</sup>, 使其在现实

应用场景中受到限制<sup>[113]</sup>. 近期研究聚焦于在解决 CTDE 范式中的缺陷的基础上提高多智能体合作效率<sup>[27-28, 116]</sup>. 基于 CTDE 范式的算法在集中训练阶段利用全局状态指导智能体, 在分散执行阶段, 每个智能体只能根据其局部观测来选择动作, 而缺乏一个共享的信号来提高智能体的合作效率<sup>[116]</sup>. 解决上述问题的有效方法之一是基于共识学习提高智能体的合作意愿. 共识学习的主要思想是所有智能体在同一时刻的观测都是相同全局状态的不同表示, 并且智能体在分散执行过程中可以利用不同的局部观察推断出对全局状态的相同共识, 进而明确地选择合作行为. 为了避免基于通信的共识学习中有限带宽与通信开销等问题, 文献 [28] 通过显式地指导智能体在分散执行过程中作出决策, 利用对比学习使得智能体在没有通信的离散空间中推断出相同的共识.

面向复杂现实世界中的合作多智能体博弈, 当前研究也注重鼓励智能体进行广泛探索、学习更适合部分可观察环境的随机策略和产生多样化的合作行为来提高多智能体合作效率<sup>[60, 66, 78]</sup>. 针对现实世界中动态环境变化会导致智能体的角色改变的情形, 文献 [78] 使用注意力机制来促使全局状态关注不同的角色表示, 根据不同角色智能体的行为和动力动态地分配子任务, 进而促进多样化的合作行为的涌现. 文献 [113] 基于熵正则化方法设计了分布式的多智能体强化学习算法, 通过在强化学习的目标中添加一个基于策略的熵正则项以激励策略利用充分随机的动作来探索环境, 同时提高样本有效性. 智能体间有效的信息传递也是提高智能体合作效率的重要手段<sup>[42]</sup>. 在合作多智能体博弈中智能体可以通过通信的方式获取其他智能体的信息, 进而缓解部分可观测问题, 增强与其他智能体的合作性能<sup>[14, 16, 117]</sup>. 基于通信的合作多智能体强化学习也是多智能体合作博弈中最受关注的研究方向, 经典研究可归纳为与谁通信、何时通信、通信内容、如何组合和整合接

收到的信息等维度, 进而促进智能体间有效合作, 共同完成合作任务<sup>[45, 51, 55, 118]</sup>. 此外, 文献 [19, 42] 梳理了基于通信的多智能体强化学习的发展历程, 并对此进行了详细的介绍.

近期有关通信的合作多智能体强化学习对原有通信机制进行改善以达到期望的博弈效果. 为了解决原有预定义的通信架构限制了智能体间的有效通信问题<sup>[14, 16]</sup>, 最新的通信机制通过设计可学习的架构来提高智能体的通信效率<sup>[90]</sup>. 图神经网络 (Graph neural network, GNN) 是基于通信的多智能体强化学习中的有效架构<sup>[55]</sup>, 但它容易受到对抗性攻击和噪声干扰的影响, 为了解决这一问题, 文献 [119] 引入图信息瓶颈优化的鲁棒通信学习机制, 该机制能够在保持通信学习有效性的同时实现鲁棒性.

此外, 研究者也基于通信的合作博弈从解决实际问题的角度出发, 考虑现实世界中的诸多限制来设计有效的通信机制. 文献 [120] 在观测和通信范围受限的情况下, 基于 Transformer 架构设计了一种可扩展的通信机制, 以解决现有方法在智能体数量变化时的不可扩展性问题. 虽然智能体只能向可观测范围内的智能体发送消息, 但通过与邮件发送类似的消息链路转发信息, 可以实现与观察范围之外的智能体合作. 表 2 对经典和最新的合作多智能体强化学习中通信机制进行了不同维度的分类. 本文发现最新的通信机制旨在解决在合作多智能体博弈中各类通信问题, 主要聚焦在通信受限条件下 (如噪声干扰<sup>[121]</sup>、通信时延<sup>[70]</sup>、信息缺失<sup>[122]</sup> 和恶意攻击<sup>[123]</sup> 等) 设计鲁棒的通信机制<sup>[119]</sup>, 也更加注重通信信息设计<sup>[124]</sup>, 关注点也由智能体间的通信扩展到人机通信<sup>[39]</sup>. 相关研究目的更加广泛, 致力于提升通信效率<sup>[90]</sup>、提高通信机制可扩展性<sup>[120]</sup> 和有效降低通信成本<sup>[79, 125]</sup> 等.

## 2.2 对抗博弈

在对抗博弈中, 以追逃问题<sup>[17-18]</sup>、德州扑克<sup>[25]</sup> 为

表 2 合作多智能体强化学习中通信机制分类

Table 2 Classification of communication mechanisms in cooperative multi-agent reinforcement learning

维度	分类	通信机制
通信约束	带宽约束	DIAL <sup>[14]</sup> , RIAL <sup>[16]</sup> , NDQ <sup>[57]</sup> , ETCNet <sup>[118]</sup> , TCOM <sup>[125]</sup>
	信息时延	DACOM <sup>[70]</sup> , RGMComm <sup>[79]</sup>
	噪声干扰	MAGI <sup>[119]</sup> , DACOM <sup>[70]</sup>
通信策略	预设定的	DIAL <sup>[14]</sup> , RIAL <sup>[16]</sup> , CommNet <sup>[45]</sup> , TarMAC <sup>[55]</sup>
	可学习的	NDQ <sup>[57]</sup> , ATOC <sup>[51]</sup> , ETCNet <sup>[118]</sup> , MAGI <sup>[119]</sup> , TEM <sup>[120]</sup> , DACOM <sup>[70]</sup> , RGMComm <sup>[79]</sup> , TCOM <sup>[125]</sup>
通信对象	所有智能体	CommNet <sup>[45]</sup> , TarMAC <sup>[55]</sup> , ETCNet <sup>[118]</sup> , DACOM <sup>[70]</sup>
	邻居智能体	MAGI <sup>[119]</sup> , RGMComm <sup>[79]</sup>
	特定智能体	ATOC <sup>[51]</sup> , TEM <sup>[120]</sup> , TCOM <sup>[125]</sup>

例, 智能体互相竞争且具有完全对立目标, 旨在最大化自己的收益或最小化损失<sup>[26]</sup>. 实际上, 在多智能体系统中存在多种对抗博弈, 文献 [15, 17–18] 等对此进行了系统性概述, 为研究者提供了宝贵的知识框架和理论基础. 本节聚焦于多智能体强化学习中的对抗博弈, 介绍相关研究的最新进展与发展趋势, 有关其他对抗博弈的详细分析可参考上述综述.

### 2.2.1 对抗博弈建模与均衡求解

在多智能体强化学习中, 常用零和马尔科夫博弈 (Zero-sum Markov game) 建模多智能体的竞争序贯决策问题<sup>[47, 126–127]</sup>, 其中各方博弈者的奖励加和为零. 根据博弈者的数量不同, 可将零和马尔科夫博弈分为双人零和马尔科夫博弈、双团队零和马尔科夫博弈和多人零和马尔科夫博弈<sup>[18]</sup>.

此外, 零和博弈<sup>[29–30]</sup>、扩展式博弈<sup>[128]</sup>、标准式博弈<sup>[129–130]</sup>、Stackelberg 博弈<sup>[131]</sup> 和零和微分博弈<sup>[132–133]</sup> 等都是多智能体系统中常见的对抗博弈模型. 上述博弈模型并不互相独立, 而是在不同层面彼此交叉<sup>[17, 128]</sup>. 例如, 零和博弈意味着博弈者的收益或损失与对方博弈者的损失或收益相对应, 总和为零; 而微分博弈是用于描述连续时间演化的博弈模型, 不仅适用于刻画对抗场景<sup>[132]</sup>, 也可用于建模混合博弈问题<sup>[134]</sup>. 零和微分博弈由于适用于描述实际追逐博弈场景而受到研究者的关注<sup>[132]</sup>. 对于零和微分博弈, 其关键是求解 HJI (Hamilton-Jacobi-Isaacs) 方程获取对应值函数, 然而应用传统解析方法求粘性解会面临非线性方程求解的困难, 而强化学习算法能处理高维非线性数据, 目前已成为求解 HJI 方程的前沿方法<sup>[7]</sup>.

不同于合作博弈以明确的奖励作为优化目标, 对抗多智能体强化学习的目标是以一种反复试验的方式学习均衡<sup>[29]</sup>, 从而获得最优决策. Nash 均衡是博弈的核心解概念, 下面给出常见的 Nash 均衡定义.

**定义 1**<sup>[129]</sup>. 在扩展式与标准式博弈中, 对任意博弈者  $i \in \mathcal{N}$ , 如果策略  $\pi_i^*$  是对其他博弈者策略  $\pi_{-i}^*$  的最佳响应, 即,

$$\pi_i^* \in \arg \max_{\pi_i \in \Phi_i} u_i(\pi_i, \pi_{-i}^*), \quad \forall \pi_i \in \Phi_i, i \in \mathcal{N} \quad (7)$$

或者

$$u_i(\pi_i^*, \pi_{-i}^*) \geq u_i(\pi_i, \pi_{-i}^*), \quad \forall \pi_i \in \Phi_i, i \in \mathcal{N} \quad (8)$$

则称策略组合  $(\pi_i^*, \pi_{-i}^*) \in \Phi$  是一个 Nash 均衡.  $\Phi_i$ ,  $u_i$  分别代表博弈者  $i$  的策略组合和效用函数,  $\Phi := \prod_{i \in \mathcal{N}} \Phi_i$  代表所有博弈者的策略组合的集合. 特别地, 如果对任意博弈者  $i \in \mathcal{N}$ ,  $u_i(\pi_i^*, \pi_{-i}^*) \geq u_i(\pi_i,$

$\pi_{-i}^*) - \epsilon$ ,  $\epsilon \geq 0$ , 则称策略组合  $(\pi_i^*, \pi_{-i}^*)$  是一个  $\epsilon$ -近似 Nash 均衡.

**定义 2**<sup>[127]</sup>. 在双人零和马尔科夫博弈中, 定义目标最大化博弈者对目标最小化博弈者的固定策略  $\nu$  给出的最佳响应为

$$V_t^{*, \nu}(s) = \max_{\pi} V_t^{\pi, \nu}(s) \quad (9)$$

对称地, 定义目标最小化博弈者对目标最大化博弈者的固定策略  $\pi$  给出的最佳响应为

$$V_t^{\pi, *}(s) = \min_{\nu} V_t^{\pi, \nu}(s) \quad (10)$$

如果存在一对策略组合  $(\pi^*, \nu^*)$  使得

$$V_1^{\pi^*, \nu^*}(s_1) = V_1^{\pi^*, *}(s_1) = V_1^{*, \nu^*}(s_1), \quad s_1 \in \mathcal{S} \quad (11)$$

则  $(\pi^*, \nu^*)$  为双人零和马尔科夫博弈的 Nash 均衡且互为对方的最佳响应.

由 Nash 均衡的定义可知, 当所有博弈者的策略组合处在一个 Nash 均衡状态时, 每个智能体都根据其他智能体的最佳响应行动, 这意味着任何一个博弈者在给定其他博弈者的最佳策略组合的情况下, 都无法通过单方面地改变自身的策略来提高各自的收益. 此外, 根据智能体学习目标与环境设定的差别, 多智能体强化学习中存在多种均衡解的概念, 如满足马尔科夫性的完美均衡<sup>[41]</sup> (Perfect Markov equilibrium, PME)、有限理性下量子响应均衡<sup>[76]</sup> (Quantal response equilibrium, QRE)、多人零和博弈中团队最大化最小化均衡<sup>[129]</sup> (Team-maxmin equilibrium, TME) 等.

针对零和马尔科夫博弈中 Nash 均衡策略的求解, 传统的多智能体强化学习算法基于 Neumann 提出的极小极大理论发展了 Minimax-Q 算法<sup>[9]</sup>. 在近期研究中, 求解零和马尔科夫博弈的多智能体强化学习算法多基于函数近似<sup>[29, 95, 126–127]</sup>、基于策略梯度<sup>[61, 71, 135]</sup> 和基于后验采样<sup>[136–138]</sup>. 基于函数近似的多智能体强化学习算法利用参数化的函数 (如神经网络) 来近似值函数或策略, 可将其分为一般函数近似<sup>[29, 95]</sup> 和线性函数近似<sup>[126–127]</sup> 的方法. 函数近似常用于具有大状态和动作空间的零和马尔科夫博弈的均衡求解. 例如, 文献 [95] 将一般函数近似的强化学习算法推广到多智能体系统中, 所设计的新算法 Golf with Exploiters 利用智能体的弱点督促其学习, 该算法适用于在大状态空间中找到 Nash 均衡策略, 且其样本复杂度与状态空间的大小无关. 研究者们也致力于将单智能体强化学习中基于策略梯度的方法扩展到多智能体强化学习中, 并证明了在双人零和马尔科夫博弈中实现 Nash 均衡的全局收敛性<sup>[135]</sup>, 为设计有效的对抗多智能体强化学习算

法提供了理论基础<sup>[30, 139]</sup>. 后验采样算法多用于解决不确定环境下强化学习问题<sup>[136]</sup>, 文献 [137] 将单智能体解耦系数的复杂性度量推广到多智能体系统中, 并在具有一般函数近似的双人零和马尔科夫博弈下, 设计了可以保证一定遗憾率的自博弈后验采样算法 (Self-play posterior sampling, SPPS).

### 2.2.2 对抗博弈的多重挑战

随着博弈对抗场景从棋牌类的分步对抗游戏转向至多人实时战略游戏、大规模无人集群对抗等现实应用的过程中, 主要存在 Nash 均衡求解难、不完美信息和非理想通信条件等挑战, 而在有限的计算资源和样本下, 高效快速算法设计的难度也是对抗博弈中的常见难题. 研究者致力于探索新的理论和方法来解决上述挑战并提升智能体的博弈对抗能力.

检验一个马尔科夫博弈中 Nash 均衡是否唯一是 NP 难问题<sup>[41]</sup>. 相比找到确切的均衡, 研究者通常更倾向于找到一个  $\epsilon$ -近似均衡<sup>[71]</sup> 或者研究比 Nash 均衡更宽松的均衡解, 如相关均衡<sup>[140]</sup> (Correlated equilibrium, CE) 和粗相关均衡<sup>[31]</sup> (Coarse correlated equilibrium, CCE). CE 和 CCE 是比 Nash 均衡更弱的稳定性概念, 并允许一定程度的策略相关性, 而不是要求每个智能体独立选择最优策略. 在状态空间和动作空间较大时, CCE 作为一种在对抗博弈中寻找 Nash 均衡的替代方法<sup>[31]</sup>, 已经引起学者们广泛关注<sup>[95, 127]</sup>.

不完美信息也是对抗博弈中的重要挑战, 是指博弈者需要在对手存在隐藏信息时作出决策<sup>[141]</sup>, 德州扑克和 Dota 2 是不完美信息博弈的典型代表<sup>[25]</sup>. 在多智能体强化学习领域, 不完美信息博弈问题可被建模为部分可观的马尔科夫博弈<sup>[81, 128]</sup> (Partially observation Markov games, POMG), 它在马尔科夫博弈的基础上引入观测函数  $\Omega(s, i)$  用于衡量智能体  $i$  从当前状态  $s \in \mathcal{S}$  观测到  $o_i \in \mathcal{O}$  的概率, 其中状态空间  $\mathcal{S}$  具有部分可观性. 虽然部分可观的马尔科夫博弈也可用于描述合作博弈场景, 但对抗场景中智能体可通过刻意隐藏己方信息的方式使得其他智能体的观测更加受限, 而合作场景中的智能体则希望互相能够建立观测上的联系. 此类博弈也常以不完美信息的扩展式博弈 (Imperfect extensive form game, IIEFG) 为模型<sup>[128]</sup>. 反事实遗憾最小化算法<sup>[142]</sup> (Counterfactual regret minimization, CFR) 和基于虚拟自博弈<sup>[25]</sup> (Fictitious self-play, FSP) 的强化学习算法是解决不完美信息博弈的主要方法, 在德州扑克、国际象棋、围棋中取得了巨大成功<sup>[17-18, 25]</sup>. 遗憾值 (Regret) 刻画了算法取得的收

益与最优策略对应收益的差值, 无悔则意味着算法的平均遗憾值趋于零. CFR 算法是博弈论范式下的先进代表性算法之一, 它多以无悔学习算法进行自博弈, 但 CFR 算法并不完全适用于环境未知的强化学习环境<sup>[138]</sup>. 另一方面, FSP 算法通过自我博弈的方式进行策略优化, 博弈者根据对手的历史行为做出最佳响应, 迭代至 Nash 均衡, 无须与人类专家进行对抗<sup>[25, 95]</sup>. CFR 算法与 FSP 算法为多智能体强化学习中解决不完美信息博弈提供了理论基础, 在求解多智能体系统对抗博弈问题中已成为通用的学习框架<sup>[128, 141]</sup>. 文献 [128] 将双人零和不完美信息扩展式博弈建模为部分可观的马尔科夫博弈, 基于 CFR 算法提出了平衡反事实遗憾最小化算法 (Balanced counterfactual regret minimization), 改进了在双人零和马尔科夫博弈中寻找  $\epsilon$ -近似 Nash 均衡的样本复杂度. 虽然有关 CFR 算法已有许多变体, 但算法程序大多都基于人工设计<sup>[17-18]</sup>. 为了减少对专家知识的依赖, AutoCFR 框架采用元学习的方法来自动设计 CFR 算法的新变体<sup>[141]</sup>, 并将其推广到不同的不完美信息博弈中. 表 3 列举了以强化学习为代表的常见求解算法, 文献 [15, 17-18] 从不同维度梳理了以对抗博弈中的多种求解方法, 相关算法的具体介绍可以参考上述文献.

表 3 对抗博弈中常见算法分类  
Table 3 Classification of common algorithms in adversarial games

算法分类	算法名称
多智能体强化学习类	基于函数近似 ONEMG <sup>[29]</sup> , OMVI-NI <sup>[126]</sup> Nash-UCRLVTR <sup>[127]</sup>
	基于策略梯度 IPG <sup>[135]</sup> , OGD <sup>[139]</sup> CoPO <sup>[61]</sup> , IPG-MAX <sup>[71]</sup>
虚拟自博弈的强化学习	FSP <sup>[25]</sup> , NFSP <sup>[148]</sup>
后验采样的强化学习	PSRL <sup>[136]</sup> , SPPS <sup>[137]</sup>
反事实遗憾最小化类	CFR <sup>[142]</sup> , AutoCFR <sup>[141]</sup>

由于受到欺诈、隐藏和低质量信息的影响, 对抗博弈主要是在非理想通信条件下进行智能体间的信息交互<sup>[20]</sup>, 给博弈者通过准确估计对方行为进行有效决策带来严峻挑战. 与合作博弈相比, 对抗博弈中不仅关注通信机制的鲁棒性<sup>[144]</sup>, 也更加注重通信的安全性<sup>[20, 145]</sup>. 大多数现有的求解分布式 Nash 均衡的算法都假设通信网络是安全的, 然而在智能电网、交通系统等实际应用中, 当通信网络遭受拒绝服务 (Denial of service, DoS) 攻击时无法维持正常的通信链路, 故不可靠通信网络下分布式 Nash 均衡寻求问题成为近期研究热点<sup>[20, 145]</sup>. 在商业竞争中博弈者之间信息交换带来的隐私问题也引起研究者

关注<sup>[146-147]</sup>, 文献 [146] 通过添加随机噪声来保护博弈者含有敏感信息的隐私. 实际上, 对抗博弈中的通信也受到多种限制. 例如博弈者只能通过通信图与相邻博弈者进行局部通信来优化自己的策略<sup>[144, 148]</sup>. 解决这一问题的有效方法是基于多轮通信设计分布式算法来收敛到全局的 Nash 均衡状态<sup>[148]</sup>. 此外, 文献 [144] 在通信资源受限和执行器故障的情况下, 基于分层思想设计了分布式算法, 利用控制层和决策层的协同优化实现了均衡状态.

## 2.3 混合博弈

混合博弈中具有合作和对抗双重动机<sup>[149]</sup>. 区别于合作博弈与对抗博弈, 混合博弈中智能体间既非完全合作也非完全对抗, 每个智能体的奖励函数各不相同. 因此, 混合博弈更加符合现实世界的复杂交互场景. 例如, 在智能电网中, 能源管理和能源市场存在互相竞争, 而在经济调度上是合作关系<sup>[4]</sup>.

### 2.3.1 混合博弈模型与学习方法

在多智能体强化学习中, 混合博弈常使用一般和马尔科夫博弈 (General-sum Markov game) 建模<sup>[33, 150]</sup>, 其中智能体间的奖励函数关系没有约束, 每个智能体最大化各自奖励<sup>[77]</sup>. 由于混合博弈涉及多方合作与对抗, 不同于对抗博弈以 Nash 均衡作为主要求解目标, 因此, 多智能体混合博弈的求解目标呈现出多样化. 帕累托最优 (Pareto optimality) 也是混合博弈中多智能体系统的求解目标之一<sup>[15]</sup>, 希望通过达到帕累托最优来实现社会最优结果.

现有一般和马尔科夫博弈均衡求解的研究主要聚焦于 Nash 均衡<sup>[89]</sup>、CE<sup>[150]</sup>、CCE<sup>[151]</sup> 和 Stackelberg 均衡<sup>[152]</sup> 这四类均衡. 除 Nash 均衡外, CE 和 CCE 是一般和马尔科夫博弈中均衡的两个标准概念, 这三种均衡关系为: Nash 均衡  $\subset$  CE  $\subset$  CCE<sup>[33]</sup>. 由于 Stackelberg 博弈均衡以领导者-追随者的分层机制为框架, 允许某些博弈者优先进行决策并强制其他博弈者执行其策略, 这符合帕累托最优的目的. 因此, 在帕累托最优意义下, Stackelberg 均衡被认为是多智能体强化学习中比 Nash 均衡更优的收敛目标<sup>[152]</sup>. 然而, 在一般和马尔科夫博弈中单个智能体常常只关注自身奖励的最大化, 导致无法达到整体最优解, 因而收敛到帕累托意义上的次优解. 为了解决这一问题, 文献 [77] 设计了基于对手意识的去中心化算法, 激励智能体产生基于互惠目标的合作行为, 解决了因忽视合作的优势而导致的社会困境问题.

在一般和马尔科夫博弈中, 寻找 Nash 均衡通常在计算上是困难的, 并属于 PPAD (多项式时间

有界) 完全问题<sup>[33]</sup>. PPAD 完全说明在多项式时间内很难求解 Nash 均衡. 求解一般和马尔科夫博弈的主要困难在于单智能体强化学习中的算法不能直接扩展到一般和马尔科夫博弈中, 而求解零和马尔科夫博弈的算法在一般和马尔科夫博弈中也无法收敛到 Nash 均衡, 为求解混合博弈带来挑战.

一般和马尔科夫博弈的目标主要分为两类, 分别是学习均衡的样本复杂性<sup>[33]</sup> 和关于任意对手的遗憾<sup>[34, 150]</sup>. 学习均衡的样本复杂性是指为了达到期望的学习效果需要多少样本, 其研究重点在于找出达到均衡时的样本数量. 在样本复杂性设置中, 通常假设所有智能体使用相同的算法进行学习. 关于任意对手的遗憾关注激励智能体在应对多种策略时最小化长期遗憾. 文献 [151, 153] 基于 V-learning 思想设计算法求解 CCE, 其算法的样本复杂性只与算法轮次  $T$  有关, 而与智能体的个数无关, 打破了传统算法中样本复杂性与智能体个数相关的限制. 在智能体只能获得有限信息的设定下, 文献 [34] 针对非平稳多智能体强化学习环境中的均衡学习问题, 设计了去中心化的算法来实现次线性的动态遗憾. 次线性遗憾意味着随着时间的推移, 遗憾的增长速度慢于线性增长, 这是长期学习过程中的一个理想特性.

传统的多智能体强化学习主要使用基于值和策略梯度的算法来解决一般和马尔科夫博弈, 但是此类算法通常需要满足较强的假设条件才可以保证收敛性. 因此, 基于策略梯度算法更适用于马尔科夫势博弈<sup>[154]</sup> (Markov potential game) 这类特殊的博弈问题. 多智能体深度确定性策略梯度 (Multi-agent deep deterministic policy gradient, MADDPG) 算法基于 AC 算法, 利用 CTDE 范式来提高学习过程的稳定性和收敛性, 是多智能体强化学习的混合任务中的经典算法<sup>[46]</sup>. 近期, 有关一般和马尔科夫博弈的研究从平稳环境的假设下<sup>[151, 155]</sup> 转移到符合现实世界的非平稳环境中<sup>[34]</sup>, 也从模型已知转移到模型未知的一般和马尔科夫博弈中<sup>[150]</sup>. 此外, 文献 [108] 发现元学习比单独学习每个任务能更快地收敛到各种博弈论解概念, 也能准确描述多智能体强化学习算法的收敛行为对初始策略的依赖性, 这促进了元学习与多智能体强化学习算法的交互融合.

### 2.3.2 面向实际问题的混合博弈

混合博弈中存在个人利益与集体利益、眼前利益与长远利益互相冲突的情况, 与人类社会、金融经济、生态学和政策制定等不同决策场景紧密相连, 在求解社会困境<sup>[77]</sup>、实现多方协作<sup>[15]</sup>、商业拍卖<sup>[110]</sup> 和外交合作<sup>[156]</sup> 中具有广泛应用.

社会困境本质上是一种决策问题,它刻画了个体在追求自身最大利益时可能会间接导致集体利益受损的现象<sup>[77]</sup>.在求解社会困境问题中,研究者通过探讨多智能体强化学习的合作和对抗行为如何随着环境而动态变化,进而揭示了智能体间具有合作的内在倾向性,并设计社会困境下提升合作水平的机制.为了解决社会困境中的冲突和矛盾,奖惩函数常被引入来促进博弈各方的合作行为.通过在智能体各自的奖励上引入内部激励函数,改变最大化个体收益的目标,可以在混合博弈场景中实现集体利益或者提升亲社会合作水平.此外,在学习过程中增加对智能体的道德和伦理约束,并通过模拟具有不同道德奖励函数的智能体交互,评估多智能体在合作、背叛或剥削行为上的不同表现和对社会带来的影响,为将道德理论应用于人工智能决策提供了新的视角<sup>[157]</sup>.

实现高效的多方协作在混合博弈中有重要意义.由于异步协作可以减少智能体间同步通信的需求,并能更好地应对通信冗余、信息延迟等现实问题,因此在近期研究中学者们也倾向于设计异步协作的策略.例如,可以利用分层的思想,即高级智能体向低级智能体发送它们的决策,进而利用博弈均衡解的概念来制定有效的智能体异步协作策略<sup>[152]</sup>.文献<sup>[158]</sup>中基于互信息感知的通信机制 How2Comm 可以在有限带宽情况下传输有价值的信息,从而在异步协作的基础上实现各智能体的高效通信.

此外,竞价拍卖<sup>[110]</sup>和无媒体外交<sup>[156]</sup>(No-press

diplomacy)也是多智能体混合博弈的主要应用场景.广告拍卖竞价博弈包含广告主和广告平台之间的合作与对抗.广告主需要在预算限制内,通过竞价机制与其他广告主以及广告平台博弈获取投放广告的机会.文献<sup>[110]</sup>提出了基于联邦学习的多智能体强化学习的广告拍卖策略,通过协调多个广告主的竞价行为,实现竞价公平性与广告平台收益的平衡,同时提升拍卖系统的竞价效率.无媒体外交博弈是一种复杂的多智能体混合博弈问题,是指多个国家在没有公开交流的情况下通过策略性的谈判、联盟、欺骗和军事行动,获得对不同地区的控制权.每个国家在考虑其他国家策略的基础上采取一系列的行动追求自身利益最大化.外交博弈要求参与博弈的国家在同一时间内提交己方行动策略.由于受到不能进行公开交流的限制,博弈者主要通过行动间接传达信息,从而增加了博弈的复杂性和战略深度<sup>[156]</sup>.

### 2.4 小结

图 2 针对合作、对抗和混合博弈中不同研究重点展示了本节中的多智能体博弈的研究框架.虽然上述三种博弈的侧重点各有不同,但解决实际需求是多智能体博弈决策技术发展的整体目标,这促进了多智能体强化学习的基线测试环境的开发和新算法的涌现.此外,多学科知识交叉融合也使得人工智能技术获得突破,有望在实际博弈场景中实现决策的高效求解.

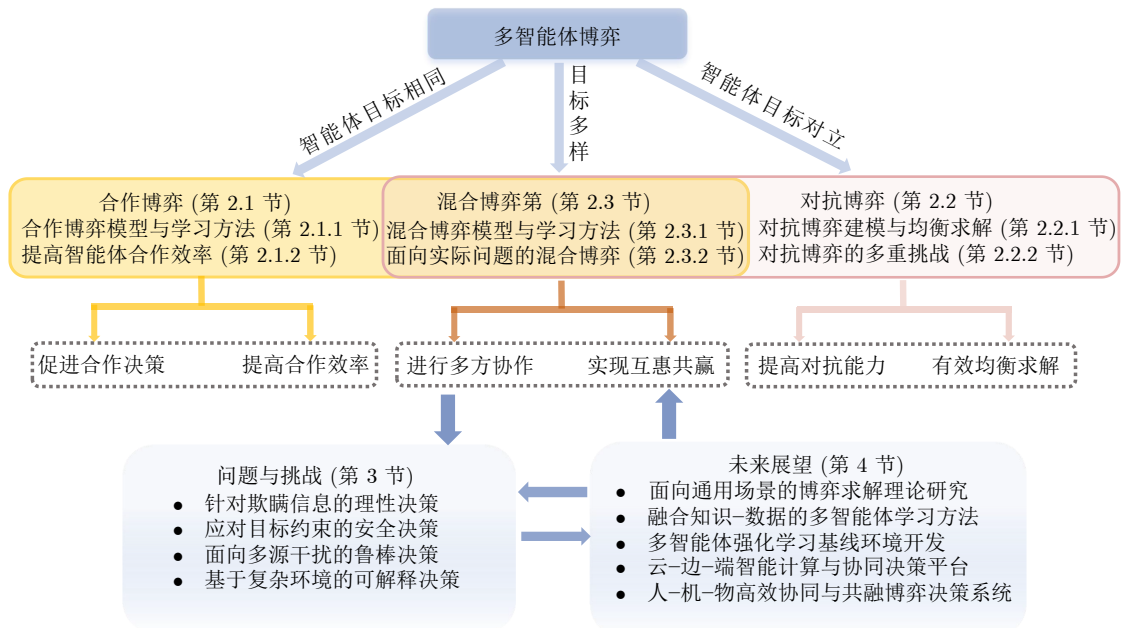


图 2 多智能体博弈研究框架  
Fig.2 Research framework of multi-agent game

### 3 应用与挑战

#### 3.1 多领域应用

随着强化学习中新技术、新算法的涌现以及与其他学习方法的交叉融合,多智能体强化学习不仅是多智能体博弈决策技术发展的重要推动力,也成为解决现实世界中各类问题的有效手段。有关多智能体强化学习的相关应用场景正逐渐从特定任务到现实世界过渡,从典型的实时战略游戏转向智能无人系统、能源系统、农业生产、生物医药和经济金融等领域发展,呈现出多样化应用趋势。例如,智能化军事中优化动态资源分配方案,实现无人机群智能调度与协同作战<sup>[2]</sup>;能源系统中优化智能电网的调度和运行<sup>[4,6]</sup>;自动驾驶中实现协同驾驶与紧急避障、智能导航系统自主智能决策<sup>[5]</sup>;机器人协作中优化路径规划与实现多目标任务分配<sup>[104]</sup>;农业生产中促进生产智能化,推动实现细粒度精准农业任务<sup>[93]</sup>;金融市场与经济管理中优化智能交易策略与设计优化投资决策,完成广告拍卖竞价<sup>[110]</sup>。

#### 3.2 问题与挑战

在多智能体系统和多智能体强化学习走向应用对象开放化、应用问题具身化、应用场景复杂化的过程中,本文从博弈求解与实际应用的角度出发,总结出如下问题与挑战。

**针对欺瞒信息的理性决策。**大多文献中均衡解的概念假设博弈者是完全理性的,虽然有少数文献提出了有限理性下均衡的新概念,然而并不适用于实际对抗场景中的多个维度强欺骗和强非理性情况。因此,针对有限理性对手或欺骗型对手的非平稳策略和虚假欺瞒信息,需要博弈者在有限理性的基础上尽可能采取最佳响应以应对虚假欺瞒信息。

**应对目标约束的安全决策。**为了满足实际安全需求,一种有效的多智能体强化学习算法是在目标函数中添加安全约束条件,然而过分考虑安全会导致求解过于保守,需要评估添加的约束条件是否会影响智能体完成实际博弈目标,并在鲁棒性和安全性之间取得平衡。因此,实现智能体期望的博弈效果与保证安全性之间的平衡是智能体安全决策设计的难点。

**面向多源干扰的鲁棒决策。**现有文献多针对单一扰动或特定约束下的鲁棒策略,有关扰动和约束的假设阻碍了将博弈决策部署到包含多源未知噪声(测量噪声、建模误差、未知扰动等)的真实场景中。虽然智能体可以通过学习多样化策略以应对多源未知干扰,但如何评估和选择最优且鲁棒的策略

对算法的计算复杂度带来挑战。

**基于复杂环境的可解释决策。**现实环境的复杂性使得智能体具有动态变化的博弈关系,研究人员致力于发展可解释的多智能体强化学习算法,它包括对环境、智能体任务和决策的解释。可解释性希望人们能够理解环境状态转移内部规律、智能体的任务目标与人工智能算法所作的决策。在复杂环境下揭示人工智能算法决策背后的逻辑和依据是可解释性决策的重要目标,并期望智能体的决策与人类的目标、道德和价值观相一致。

### 4 未来展望

作为新一代人工智能的热点研究领域,多智能体博弈决策的理论和应用发展方兴未艾,以大模型为代表的人工智能新浪潮将多智能体博弈决策的研究推向新的高度,也使得多智能体强化学习具有更广阔、更现实的应用场景。

#### 4.1 面向通用场景的博弈求解理论研究

多智能体博弈朝着多种场景扩展、迁移等方向不断发展使得智能体的理性程度、认知层次与决策能力复杂多变,这对多智能体博弈模型从仿真环境到实际环境的转换提出了挑战,并要求博弈求解算法在未知场景中具有良好的探索能力。通过简化模型来设计多智能体博弈求解算法会严重降低模型刻画实际博弈问题的准确性和有效性,因此,需要研发面向通用博弈决策大模型,研究能保障多项博弈决策任务的均衡收敛性和满足强鲁棒、高安全、可泛化的博弈求解理论方法。

#### 4.2 融合知识-数据的多智能体学习方法

随着多学科交叉融合趋势的日益显著,融合多学科知识是实现高效多智能体强化学习方法的重要途径。例如,对于具有自身动力学的智能体来说,全面刻画其运动学和动力学行为,充分考虑其物理特性约束,对于实现多智能体学习算法的应用具有重要意义。因此,通过基于机理知识并融合多模态大数据,研发有具身特性的多智能体学习算法(多体具身智能),以及通过智能体之间在语义信息、特征信息和测量数据中的协作交互,增强模型自适应力,提高算法可扩展性和收敛性能,将推动多智能体强化学习应用于更广阔的真实世界。

#### 4.3 多智能体强化学习基线环境开发

人工智能的发展促进了多智能体强化学习算法的创新。随着新算法的不断涌现,需要开发新的基线测试平台来适应和评估新算法在现实场景中的表

现,同时满足多智能体博弈模型中高维状态空间和连续动作空间的巨大算力需求.设计细粒度基线环境、搭建普适性仿真平台来支持博弈决策算法研究,通过元宇宙思想解决任务场景碎片化、多元信息异构化的问题,实现算法可视化是推进多智能体强化学习算法从仿真平台应用到真实世界的关键途径.

#### 4.4 云-边-端智能计算与协同决策平台

人工智能大模型对大量训练数据、计算资源和强大算力的迫切需求推动了云计算的创新与发展,结合云上大规模计算优势并利用端侧态势感知能力,使得云上大模型与端侧小模型计算平台的协同成为可能,进而实现云上、端侧和端云边的计算资源的高效利用.针对云-边-端协同框架面临的应用瓶颈,利用因果强化学习克服云上预训练大模型中数据偏差问题,基于迁移学习方法解决端侧有限数据而导致的小样本学习挑战,构建端侧个性化的多智能体学习算法,并研发专用芯片进而构建云-边-端智能计算与协同决策平台,提升个性化、高效化能力,将对多体具身智能的应用具有重要意义.

#### 4.5 人-机-物高效协同与共融博弈决策系统

新一代人工智能技术的发展正在推动人、机、物三元空间的有机融合,人工智能与大数据、通信技术的交叉融合极大地促进了社会空间、信息空间和物理空间中人类、机器、对象三者的高效协作.人机关系是人、机、物协同的关键所在,基于人工智能的认知推理、自主学习、智能决策的优势,利用人类的主观感知、态势理解、判断指导与反馈优化的能力,通过人类智能和机器智能的交互演进,突破小样本、可解释性差、算力稀缺等瓶颈,搭建人-机-物的高效协同与共融博弈决策系统至关重要.

## 5 总结

本文聚焦于合作、对抗以及混合三种多智能体强化学习任务下的最新研究进展,对基于多智能体强化学习的博弈进行了系统性综述.针对智能体不同的学习目标对多智能体博弈模型、学习方法进行了梳理分类,考虑三种博弈模式中智能体交互决策的不同侧重问题,探讨了如何提高智能体间合作效率、提高智能体对抗能力的学习方法,分析了实际场景中混合博弈的求解方法.最后,总结了多智能体博弈决策的现阶段挑战,并从顺应人工智能发展新浪潮的角度出发,展望了多智能体强化学习和博弈决策中理论、技术与实际应用的发展愿景,希望能为该领域的研究者提供有意义的参考.

## References

- Miao Qing-Hai, Wang Xing-Xia, Yang Jing, Zhao Yong, Wang Yu-Tong, Chen Yuan-Yuan, et al. From foundation intelligence to general intelligence: The state-of-art and perspectives of GenAI and AGI based on foundation models. *Acta Automatica Sinica*, 2024, **50**(4): 674-687  
(缪青海, 王兴霞, 杨静, 赵勇, 王雨桐, 陈圆圆, 等. 从基础智能到通用智能: 基于大模型的 GenAI 和 AGI 之现状与展望. *自动化学报*, 2024, **50**(4): 674-687)
- Shi Wei, Feng Yang-He, Cheng Guang-Quan, Huang Hong-Lan, Huang Jin-Cai, Liu Zhong, et al. Research on multi-aircraft cooperative air combat method based on deep reinforcement learning. *Acta Automatica Sinica*, 2021, **47**(7): 1610-1623  
(施伟, 冯阳赫, 程光权, 黄红蓝, 黄金才, 刘忠, 等. 基于深度强化学习的多机协同空战方法研究. *自动化学报*, 2021, **47**(7): 1610-1623)
- Liu Hua-Ping, Guo Di, Sun Fu-Chun, Zhang Xin-Yu. Morphology-based embodied intelligence: Historical retrospect and research progress. *Acta Automatica Sinica*, 2023, **49**(6): 1131-1154  
(刘华平, 郭迪, 孙富春, 张新钰. 基于形态的具身智能研究: 历史回顾与前沿进展. *自动化学报*, 2023, **49**(6): 1131-1154)
- Xiong Luo-Lin, Mao Shuai, Tang Yang, Meng Ke, Dong Zhao-Yang, Qian Feng. Reinforcement learning based integrated energy system management: A survey. *Acta Automatica Sinica*, 2021, **47**(10): 2321-2340  
(熊珞琳, 毛帅, 唐漾, 孟科, 董朝阳, 钱锋. 基于强化学习的综合能源系统管理综述. *自动化学报*, 2021, **47**(10): 2321-2340)
- Kiran B R, Sobh I, Talpaert V, Mannion P, Al Sallab A A, Yogamani S, et al. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2022, **23**(6): 4909-4926
- Qu Y, Ma J M, Wu F. Safety constrained multi-agent reinforcement learning for active voltage control. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence. Jeju, South Korea: IJCAI, 2024. 184-192
- Wang Long, Huang Feng. An interdisciplinary survey of multi-agent games, learning, and control. *Acta Automatica Sinica*, 2023, **49**(3): 580-613  
(王龙, 黄锋. 多智能体博弈、学习与控制. *自动化学报*, 2023, **49**(3): 580-613)
- Wen M N, Kuba G J, Lin R J, Zhang W N, Wen Y, Wang J, et al. Multi-agent reinforcement learning is a sequence modeling problem. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1201
- Littman M L. Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of the 11th International Conference on Machine Learning. New Brunswick, USA: Morgan Kaufmann Publishers Inc., 1994. 157-163
- Sivagnanam A, Pettet A, Lee H, Mukhopadhyay A, Dubey A, Laszka A. Multi-agent reinforcement learning with hierarchical coordination for emergency responder stationing. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR, 2024. 45813-45834
- Puterman M L. Markov decision processes. *Handbooks in Operations Research and Management Science*. Amsterdam: Elsevier, 1990. 331-434
- Pu Zhi-Qiang, Yi Jian-Qiang, Liu Zhen, Qiu Teng-Hai, Sun Jin-Lin, Li Fei-Mo. Knowledge-based and data-driven integrating methodologies for collective intelligence decision making: A survey. *Acta Automatica Sinica*, 2022, **48**(3): 627-643  
(蒲志强, 易建强, 刘振, 丘腾海, 孙金林, 李非墨. 知识和数据协同驱动的群体智能决策方法研究综述. *自动化学报*, 2022, **48**(3): 627-643)
- Wen M N, Wan Z Y, Zhang W N, Wang J, Wen Y. Reinforcing language agents via policy optimization with action decomposition. In: Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada:

- NeurIPS, 2024.
- 14 Foerster J N, Assael Y M, de Freitas N, Whiteson S. Learning to communicate with deep multi-agent reinforcement learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc., 2016. 2145–2153
  - 15 Hao Jian-Ye, Shao Kun, Li Kai, Li Dong, Mao Hang-Yu, Hu Shu-Yue, et al. Research and applications of game intelligence. *Scientia Sinica Informationis*, 2023, **53**(10): 1892–1923 (郝建业, 邵坤, 李凯, 李栋, 毛航宇, 胡舒悦, 等. 博弈智能的研究与应用. 中国科学: 信息科学, 2023, **53**(10): 1892–1923)
  - 16 Hausknecht M, Stone P. Deep recurrent Q-learning for partially observable MDPs. In: Proceedings of the AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents. Arlington, USA: AAAI, 2015. 29–37
  - 17 Li X X, Meng M, Hong Y G, Chen J. A survey of decision making in adversarial games. *Science China Information Sciences*, 2024, **67**(4): Article No. 141201
  - 18 Qin R J, Yu Y. Learning in games: A systematic review. *Science China Information Sciences*, 2024, **67**(7): Article No. 171101
  - 19 Zhu C X, Dastani M, Wang S H. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-agent Systems*, 2024, **38**(1): Article No. 4
  - 20 Meng Q, Nian X H, Chen Y, Chen Z. Attack-resilient distributed Nash equilibrium seeking of uncertain multiagent systems over unreliable communication networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(5): 6365–6379
  - 21 Lechner M, Yin L H, Seyde T, Wang T H, Xiao W, Hasani R, et al. Gigastep-one billion steps per second multi-agent reinforcement learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 9
  - 22 Qu Y, Wang B Y, Shao J Z, Jiang Y H, Chen C, Ye Z B, et al. Hokoff: Real game dataset from honor of kings and its offline reinforcement learning benchmarks. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 974
  - 23 Mazzaglia P, Verbelen T, Dhoedt B, Courville A, Rajeswar S. Multimodal foundation world models for generalist embodied agents. In: Proceedings of the ICML Workshop: Multi-modal Foundation Model Meets Embodied AI. Vienna, Austria: ICML, 2024.
  - 24 Li M L, Zhao S Y, Wang Q N, Wang K R, Zhou Y, Srivastava S, et al. Embodied agent interface: Benchmarking LLMs for embodied decision making. In: Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2024.
  - 25 Heinrich J, Lanctot M, Silver D. Fictitious self-play in extensive-form games. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR, 2015. 805–813
  - 26 Wang J R, Hong Y T, Wang J L, Xu J P, Tang Y, Han Q L, et al. Cooperative and competitive multi-agent systems: From optimization to games. *IEEE/CAA Journal of Automatica Sinica*, 2022, **9**(5): 763–783
  - 27 Yu Wen-Wu, Yang Xiao-Ya, Li Hai-Chang, Wang Rui, Hu Xiao-Hui. Attentional intention and communication for multi-agent learning. *Acta Automatica Sinica*, 2023, **49**(11): 2311–2325 (俞文武, 杨晓亚, 李海昌, 王瑞, 胡晓惠. 面向多智能体协作的注意力意图与交流学习方法. 自动化学报, 2023, **49**(11): 2311–2325)
  - 28 Xu Z W, Zhang B, Li D P, Zhang Z R, Zhou G C, Chen H, et al. Consensus learning for cooperative multi-agent reinforcement learning. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI, 2023. 11726–11734
  - 29 Huang B H, Lee J, Wang Z R, Yang Z R. Towards general function approximation in zero-sum Markov games. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR, 2022.
  - 30 Alacaoglu A, Viano L, He N, Cevher V. A natural actor-critic framework for zero-sum Markov games. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 307–366
  - 31 Li C J, Zhou D R, Gu Q Q, Jordan M I. Learning two-player mixture Markov games: Kernel function approximation and correlated equilibrium. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 2410
  - 32 Li S H, Wu Y, Cui X Y, Dong H H, Fang F, Russell S. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI, 2019. 4213–4220
  - 33 Song Z A, Mei S, Bai Y. When can we learn general-sum Markov games with a large number of players sample-efficiently? In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR, 2022.
  - 34 Jiang H Z, Cui Q W, Xiong Z H, Fazel M, Du S S. A black-box approach for non-stationary multi-agent reinforcement learning. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR, 2024.
  - 35 Silver D, Huang A, Maddison C J, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, **529**(7587): 484–489
  - 36 Jaderberg M, Czarnecki W M, Dunning I, Marris L, Lever G, Castañeda A G, et al. Human-level performance in 3D multi-player games with population-based reinforcement learning. *Science*, 2019, **364**(6443): 859–865
  - 37 Samvelyan M, Rashid T, de Witt C S, Farquhar G, Nardelli N, Rudner T G J, et al. The StarCraft multi-agent challenge. In: Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems. Montreal, Canada: AAMAS, 2019. 2186–2188
  - 38 de Witt C S, Gupta T, Makoviychuk D, Makoviychuk V, Torr P H S, Sun M F, et al. Is independent learning all you need in the StarCraft multi-agent challenge? arXiv: 2011.09533, 2020.
  - 39 Seraj E, Xiong J, Schrum M, Gombolay M. Mixed-initiative multiagent apprenticeship learning for human training of robot teams. In: Proceedings of the 37th Conference on Neural Information Processing Systems. New Orleans, USA: NeurIPS, 2023.
  - 40 Wen Guang-Hui, Yang Tao, Zhou Jia-Ling, Fu Jun-Jie, Xu Lei. Reinforcement learning and adaptive/approximate dynamic programming: A survey from theory to applications in multi-agent systems. *Control and Decision*, 2023, **38**(5): 1200–1230 (温广辉, 杨涛, 周佳玲, 付俊杰, 徐磊. 强化学习与自适应动态规划: 从基础理论到多智能体系统中的应用进展综述. 控制与决策, 2023, **38**(5): 1200–1230)
  - 41 Yang Y D, Wang J. An overview of multi-agent reinforcement learning from game theoretical perspective. arXiv preprint arXiv: 2011.00583, 2020.
  - 42 Wang Han, Yu Yang, Jiang Yuan. Review of the progress of communication-based multi-agent reinforcement learning. *Scientia Sinica Informationis*, 2022, **52**(5): 742–764 (王涵, 俞扬, 姜远. 基于通信的多智能体强化学习进展综述. 中国科学: 信息科学, 2022, **52**(5): 742–764)
  - 43 Wang Xue-Song, Wang Rong-Rong, Cheng Yu-Hu. Safe reinforcement learning: A survey. *Acta Automatica Sinica*, 2023, **49**(9): 1813–1835 (王雪松, 王荣荣, 程玉虎. 安全强化学习综述. 自动化学报, 2023, **49**(9): 1813–1835)

- 44 Sun Yue-Wen, Liu Wen-Zhang, Sun Chang-Yin. Causality in reinforcement learning control: The state of the art and prospects. *Acta Automatica Sinica*, 2023, **49**(3): 661–677 (孙悦雯, 柳文章, 孙长银. 基于因果建模的强化学习控制: 现状及展望. *自动化学报*, 2023, **49**(3): 661–677)
- 45 Sukhbaatar S, Szlam A, Fergus R. Learning multiagent communication with backpropagation. In: *Proceedings of 30th International Conference on Neural Information Processing Systems*. Barcelona, Spain: Curran Associates Inc., 2016. 2252–2260
- 46 Lowe R, Wu Y, Tamar A, Harb J, Abbeel P, Mordatch I. Multi-agent actor-critic for mixed cooperative-competitive environments. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, USA: Curran Associates Inc., 2017. 6382–6393
- 47 Peng P, Wen Y, Yang Y D, Yuan Q, Tang Z K, Tang H T, et al. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play StarCraft combat games. arXiv: 1703.10069, 2017.
- 48 Peng B, Rashid T, Schroeder de Witt C A, Kamienny P A, Torr P H S, Böhm W, et al. FACMAC: Factored multi-agent centralised policy gradients. In: *Proceedings of the 35th Conference on Neural Information Processing Systems*. Virtual Event: NeurIPS, 2021.
- 49 Sunehag P, Lever G, Gruslys A, Czarnecki W M, Zambaldi V, Jaderberg M, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. In: *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*. Stockholm, Sweden: AAMAS, 2018. 2085–2087
- 50 Rashid T, Samvelyan M, Schroeder C, Farquhar G, Foerster J, Whiteson S. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In: *Proceedings of the 35th International Conference on Machine Learning*. Stockholm, Sweden: PMLR, 2018. 4295–4304
- 51 Jiang J C, Lu Z Q. Learning attentional communication for multi-agent cooperation. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Montréal, Canada: Curran Associates Inc., 2018. 7265–7275
- 52 Zheng L M, Yang J C, Cai H, Zhou M, Zhang W N, Wang Jun, et al. MAgent: A many-agent reinforcement learning platform for artificial collective intelligence. In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. New Orleans, USA: AAAI, 2018. 8222–8223
- 53 Resnick C, Eldridge W, Ha D, Britz D, Foerster J, Togelius J, et al. Pommerman: A multi-agent playground. arXiv preprint arXiv: 1809.07124, 2018.
- 54 Iqbal S, Sha F. Actor-attention-critic for multi-agent reinforcement learning. In: *Proceedings of the 36th International Conference on Machine Learning*. Long Beach, USA: PMLR, 2019. 2961–2970
- 55 Das A, Gervet T, Romof J, Batra D, Parikh D, Rabbat M, et al. TarMAC: Targeted multi-agent communication. In: *Proceedings of the 36th International Conference on Machine Learning*. Long Beach, California: PMLR, 2019. 1538–1546
- 56 Suarez J, Du Y L, Isola P, Mordatch I. Neural MMO: A massively multiagent game environment for training and evaluating intelligent agents. arXiv: 1903.00784, 2019.
- 57 Zheng C Y, Wang J H, Zhang C J, Wang T H. Learning nearly decomposable value functions via communication minimization. In: *Proceedings of the 8th International Conference on Learning Representations*. Addis Ababa, Ethiopia: ICLR, 2020.
- 58 Kurach K, Raichuk A, Stańczyk P, Zajac M, Bachem O, Espeholt L, et al. Google research football: A novel reinforcement learning environment. In: *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. New York, USA: AAAI, 2020. 4501–4510
- 59 Zhou M, Luo J, Vilella J, Yang Y D, Rusu D, Miao J Y, et al. SMARTS: Scalable multi-agent reinforcement learning training school for autonomous driving. arXiv preprint arXiv: 2010.09776, 2020.
- 60 Su J Y, Adams S, Beling P. Value-decomposition multi-agent actor-critics. In: *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. Virtual Event: AAAI, 2021. 11352–11360
- 61 Prajapat M, Azizzadenesheli K, Liniger A, Yue Y S, Anandkumar A. Competitive policy optimization. In: *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*. Virtual Event: PMLR, 2021. 64–74
- 62 Leibo J Z, Duéñez-Guzmán E A, Vezhnevets A, Agapiou J P, Sunehag P, Koster R, et al. Scalable evaluation of multi-agent reinforcement learning with melting pot. In: *Proceedings of the 38th International Conference on Machine Learning*. Virtual Event: PMLR, 2021. 6187–6199
- 63 Hong Y T, Jin Y C, Tang Y. Rethinking individual global max in cooperative multi-agent reinforcement learning. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. New Orleans, USA: Curran Associates Inc., 2022. Article No. 2350
- 64 Shao J Z, Lou Z Q, Zhang H C, Jiang Y H, He S C, Ji X Y. Self-organized group for cooperative multi-agent reinforcement learning. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. New Orleans, USA: Curran Associates Inc., 2022. 5711–5723
- 65 Jiang J C, Lu Z Q. I2Q: A fully decentralized Q-learning algorithm. In: *Proceedings of the 36th Conference on Neural Information Processing Systems*. New Orleans, USA: NeurIPS, 2022. 20469–20481
- 66 Su K F, Lu Z Q. Divergence-regularized multi-agent actor-critic. In: *Proceedings of the 39th International Conference on Machine Learning*. Baltimore, USA: PMLR, 2022. 20580–20603
- 67 Yang M Y, Zhao J, Hu X H, Zhou W G, Zhu J C, Li H Q. LDSA: Learning dynamic subtask assignment in cooperative multi-agent reinforcement learning. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. New Orleans, USA: Curran Associates Inc., 2022. Article No. 124
- 68 Pan X H, Liu M, Zhong F W, Yang Y D, Zhu S C, Wang Y Z. MATE: Benchmarking multi-agent reinforcement learning in distributed target coverage control. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. New Orleans, USA: Curran Associates Inc., 2022. Article No. 2021
- 69 Li J H, Kuang K, Wang B X, Li X C, Wu F, Xiao J, et al. Two heads are better than one: A simple exploration framework for efficient multi-agent reinforcement learning. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. New Orleans, USA: Curran Associates Inc., 2023. Article No. 878
- 70 Yuan T T, Chung H M, Yuan J, Fu X M. DACOM: Learning delay-aware communication for multi-agent reinforcement learning. In: *Proceedings of the 37th AAAI Conference on Artificial Intelligence*. Washington, USA: AAAI, 2023. 11763–11771
- 71 Kalogiannis F, Anagnostides I, Panageas I, Vlatakis-Gkaragkounis E V, Chatziafratis V, Stavroulakis S. Efficiently computing Nash equilibria in adversarial team Markov games. In: *Proceedings of the 11th International Conference on Learning Representations*. Kigali, Rwanda: ICLR, 2023.
- 72 Leroy P, Morato P G, Pisane J, Kolios A, Ernst D. IMP-MARL: A suite of environments for large-scale infrastructure management planning via MARL. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. New Orleans, USA: Curran Associates Inc., 2023. Article No. 2329
- 73 Suárez J, Isola P, Choe K W, Bloomin D, Li H X, Pinnaparaju N, et al. Neural MMO 2.0: A massively multi-task addition to massively multi-agent learning. In: *Proceedings of the 37th International Conference on Neural Information Pro-*

- cessing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 2178
- 74 Ellis B, Cook J, Moalla S, Samvelyan M, Sun M F, Mahajan A, et al. SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 1634
- 75 Liu C X, Liu G Z. JointPPO: Diving deeper into the effectiveness of PPO in multi-agent reinforcement learning. arXiv: 2404.11831, 2024.
- 76 Liu J R, Zhong Y F, Hu S Y, Fu H B, Fu Q, Chang X J, et al. Maximum entropy heterogeneous-agent reinforcement learning. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR, 2024.
- 77 Aghajohari M, Duque J A, Cooijmans T, Courville A. LOQA: Learning with opponent Q-Learning awareness. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR, 2024.
- 78 Hu Z C, Zhang Z Z, Li H X, Chen C L, Ding H Y, Wang Z. Attention-guided contrastive role representations for multi-agent reinforcement learning. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR, 2024.
- 79 Chen J D, Lan T, Joe-Wong C. RGMComm: Return gap minimization via discrete communications in multi-agent reinforcement learning. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024. 17327–17336
- 80 Geng M H, Pateria S, Subagdja B, Tan A H. Benchmarking MARL on long horizon sequential multi-objective tasks. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems. Auckland, New Zealand: AAMAS, 2024. 2279–2281
- 81 Li W Z, Ding Z H, Karten S, Jin C. FightLadder: A benchmark for competitive multi-agent reinforcement learning. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR, 2024. 27653–27674
- 82 Zhu S H, Zhou J C, Chen A J, Bai M M, Chen J M, Xu J M. MAexp: A generic platform for RL-based multi-agent exploration. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024. 5155–5161
- 83 Wang Y T, Duhan T, Li J Y, Sartoretti G. LNS2+RL: Combining multi-agent reinforcement learning with large neighborhood search in multi-agent path finding. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI, 2025.
- 84 Watkins C J C H, Dayan P. Q-learning. *Machine Learning*, 1992, 8(3–4): 279–292
- 85 Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller M. Deterministic policy gradient algorithms. In: Proceedings of the 31st International Conference on Machine Learning. Beijing, China: JMLR, 2014. I-387–I-395
- 86 Konda V R, Tsitsiklis J N. Actor-critic algorithms. In: Proceedings of the 12th International Conference on Neural Information Processing Systems. Denver, USA: MIT Press, 1999. 1008–1014
- 87 McClellan J, Haghani N, Winder J, Huang F R, Tokekar P. Boosting sample efficiency and generalization in multi-agent reinforcement learning via equivariance. In: Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2024.
- 88 Shapley L S. Stochastic games. *Proceedings of the National Academy of Sciences*, 1953, 39(10): 1095–1100
- 89 Hu J L, Wellman M P. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 2003, 4: 1039–1069
- 90 Hu S C, Shen L, Zhang Y, Tao D C. Learning multi-agent communication from graph modeling perspective. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR, 2024.
- 91 Wen G H, Fu J J, Dai P C, Zhou J L. DTDE: A new cooperative multi-agent reinforcement learning framework. *The Innovation*, 2021, 2(4): Article No. 100162
- 92 Foerster J N, Farquhar G, Afouras T, Nardelli N, Whiteson S. Counterfactual multi-agent policy gradients. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI, 2018. Article No. 363
- 93 Li W H, Wang X F, Jin B, Luo D J, Zha H Y. Structured cooperative reinforcement learning with time-varying composite action space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(11): 8618–8634
- 94 Tan M. Multi-agent reinforcement learning: Independent versus cooperative agents. In: Proceedings of the 10th International Conference on Machine Learning. Amherst, USA: Morgan Kaufmann Publishers Inc., 1993. 330–337
- 95 Jin C, Liu Q H, Yu T C. The power of exploiter: Provable multi-agent RL in large state spaces. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 10251–10279
- 96 Jiang H B, Ding Z L, Lu Z Q. Settling decentralized multi-agent coordinated exploration by novelty sharing. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024. 17444–17452
- 97 Hsu H L, Wang W X, Pajic M, Xu P. Randomized exploration in cooperative multi-agent reinforcement learning. arXiv: 2404.10728, 2024.
- 98 Shin W, Kim Y. Guide to control: Offline hierarchical reinforcement learning using subgoal generation for long-horizon and sparse-reward tasks. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2023. 4217–4225
- 99 Wang J H, Zhang Y, Kim T K, Gu Y J. Shapley Q-value: A local reward approach to solve global reward games. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 7285–7292
- 100 Wang X S, Xu H R, Zheng Y N, Zhan X Y. Offline multi-agent reinforcement learning with implicit global-to-local value regularization. In: Proceedings of the 37th Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 2282
- 101 Boggess K, Kraus S, Feng L. Explainable multi-agent reinforcement learning for temporal queries. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2023. 55–63
- 102 Zhang H, Chen H G, Boning D S, Hsieh C J. Robust reinforcement learning on state observations with learned optimal adversary. In: Proceedings of the 9th International Conference on Learning Representations. Virtual Event: ICLR, 2021.
- 103 Li Z, Wellman M P. A meta-game evaluation framework for deep multiagent reinforcement learning. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence. Jeju, South Korea: IJCAI, 2024. 148–156
- 104 Yang Z X, Jin H M, Ding R, You H Y, Fan G Y, Wang X B, et al. DeCOM: Decomposed policy for constrained cooperative multi-agent reinforcement learning. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI, 2023. 10861–10870
- 105 García J, Fernández F. A comprehensive survey on safe reinforcement learning. *The Journal of Machine Learning Research*, 2015, 16(1): 1437–1480
- 106 Chen Z Y, Zhou Y, Huang H. On the duality gap of constrained cooperative multi-agent reinforcement learning. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR, 2024.

- 107 Bukharin A, Li Y, Yu Y, Zhang Q R, Chen Z H, Zuo S M, et al. Robust multi-agent reinforcement learning via adversarial regularization: Theoretical foundation and stable algorithms. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 2979
- 108 Mao W C, Qiu H R, Wang C, Franke H, Kalbarczyk Z, Iyer R K, et al. Multi-agent meta-reinforcement learning: Sharper convergence rates with task similarity. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 2906
- 109 Tan Xiao-Yang, Zhang Zhe. Review on meta reinforcement learning. *Journal of Nanjing University of Aeronautics Astronautics*, 2021, **53**(5): 653–663  
(谭晓阳, 张哲. 元强化学习综述. *南京航空航天大学学报*, 2021, **53**(5): 653–663)
- 110 Tang X L, Yu H. Competitive-cooperative multi-agent reinforcement learning for auction-based federated learning. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2023. 4262–4270
- 111 Cao K, Xie L H. Trust-region inverse reinforcement learning. *IEEE Transactions on Automatic Control*, 2024, **69**(2): 1037–1044
- 112 Deng Y, Wang Z R, Zhang Y. Improving multi-agent reinforcement learning with stable prefix policy. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence. Jeju, South Korea: IJCAI, 2024. 49–57
- 113 Hu Y F, Fu J J, Wen G H, Lv Y Z, Ren W. Distributed entropy-regularized multi-agent reinforcement learning with policy consensus. *Automatica*, 2024, **164**: Article No. 111652
- 114 Kok J R, Vlassis N. Collaborative multiagent reinforcement learning by payoff propagation. *Journal of Machine Learning Research*, 2006, **7**: 1789–1828
- 115 Son K, Kim D, Kang W J, Hostallero D, Yi Y. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019. 5887–5896
- 116 Li W H, Wang X F, Jin B, Sheng J J, Hua Y, Zha H Y. Structured diversification emergence via reinforced organization control and hierarchical consensus learning. In: Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems. Virtual Event: AAMAS, 2021. 773–781
- 117 Mnih V, Badia A P, Mirza M, Graves A, Harley T, Lillicrap T P, et al. Asynchronous methods for deep reinforcement learning. In: Proceedings of the 33rd International Conference on Machine Learning. New York, USA: JMLR, 2016. 1928–1937
- 118 Hu G Z, Zhu Y H, Zhao D B, Zhao M C, Hao J Y. Event-triggered communication network with limited-bandwidth constraint for multi-agent reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(8): 3966–3978
- 119 Ding S F, Du W, Ding L, Zhang J, Guo L L, An B. Robust multi-agent communication with graph information bottleneck optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(5): 3096–3107
- 120 Guo X D, Shi D M, Fan W H. Scalable communication for multi-agent reinforcement learning via transformer-based email mechanism. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2023. 126–134
- 121 Tang Y J, Ren Z L, Li N. Zeroth-order feedback optimization for cooperative multi-agent systems. *Automatica*, 2023, **148**: Article No. 110741
- 122 Rachmut B, Nelke S A, Zivan R. Asynchronous communication aware multi-agent task allocation. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2023. Article No. 30
- 123 Yu L B, Qiu Y B, Yao Q M, Shen Y, Zhang X D, Wang J. Robust communicative multi-agent reinforcement learning with active defense. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024. 17575–17582
- 124 Lin Y, Li W H, Zha H Y, Wang B X. Information design in multi-agent reinforcement learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 1113
- 125 Wang X C, Yang L, Chen Y Z, Liu X T, Hajjesmaili M, Towsley D, et al. Achieving near-optimal individual regret & low communications in multi-agent bandits. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR, 2023.
- 126 Xie Q M, Chen Y D, Wang Z R, Yang Z R. Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium. In: Proceedings of the 33rd Conference on Learning Theory. Graz, Austria: PMLR, 2020. 3674–3682
- 127 Chen Z X, Zhou D R, Gu Q Q. Almost optimal algorithms for two-player zero-sum linear mixture Markov games. In: Proceedings of the 33rd International Conference on Algorithmic Learning Theory. Paris, France: PMLR, 2022. 227–261
- 128 Bai Y, Jin C, Mei S, Yu T C. Near-optimal learning of extensive-form games with imperfect information. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 1337–1382
- 129 Zhang Y Z, An B. Converging to team-maxmin equilibria in zero-sum multiplayer games. In: Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria: JMLR, 2020. Article No. 1023
- 130 Chen Y Q, Mao H Y, Mao J X, Wu S G, Zhang T L, Zhang B, et al. PTDE: Personalized training with distilled execution for multi-agent reinforcement learning. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence. Jeju, South Korea: IJCAI, 2024. 31–39
- 131 Zheng L Y, Fiez T, Alumbaugh Z, Chasnov B, Ratliff L J. Stackelberg actor-critic: Game-theoretic reinforcement learning algorithms. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI, 2022. 9217–9224
- 132 Wang J L, Jin X, Tang Y. Optimal strategy analysis for adversarial differential games. *Electronic Research Archive*, 2022, **30**(10): 3692–3710
- 133 Plaksin A, Kalev V. Zero-sum positional differential games as a framework for robust reinforcement learning: Deep Q-learning approach. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR, 2024. 40869–40885
- 134 Wang J L, Zhou Z, Jin X, Mao S, Tang Y. Matching-based capture-the-flag games for multiagent systems. *IEEE Transactions on Cognitive and Developmental Systems*, 2024, **16**(3): 993–1005
- 135 Daskalakis C, Foster D J, Golowich N. Independent policy gradient methods for competitive reinforcement learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 464
- 136 Ian O, Benjamin V R, Daniel R. (More) Efficient reinforcement learning via posterior sampling. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc., 2013. 3003–3011
- 137 Xiong W, Zhong H, Shi C S, Shen C, Zhang T. A self-play posterior sampling algorithm for zero-sum Markov games. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 24496–24523

- 138 Zhou Y C, Li J L, Zhu J. Posterior sampling for multi-agent reinforcement learning: Solving extensive games with imperfect information. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: ICLR, 2020.
- 139 Wei C Y, Lee C W, Zhang M X, Luo H P. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. In: Proceedings of the 34th Annual Conference on Learning Theory. Boulder, USA: PMLR, 2021. 4259–4299
- 140 Wang X R, Yang C, Li S X, Li P D, Huang X, Chan H, et al. Reinforcement Nash equilibrium solver. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence. Jeju, South Korea: IJCAI, 2024. 265–273
- 141 Xu H, Li K, Fu H B, Fu Q, Xing J L. AutoCFR: Learning to design counterfactual regret minimization algorithms. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2022. 5244–5251
- 142 Zinkevich M, Johanson M, Bowling M, Piccione C. Regret minimization in games with incomplete information. In: Proceedings of the 20th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2007. 1729–1736
- 143 Heinrich J, Silver D. Deep reinforcement learning from self-play in imperfect-information games. arXiv: 1603.01121, 2016.
- 144 Wang H, Luo H, Jiang Y C, Kaynak O. A performance recovery approach for multiagent systems with actuator faults in noncooperative games. *IEEE Transactions on Industrial Informatics*, 2024, **20**(5): 7853–7861
- 145 Zhong Y F, Yuan Y, Yuan H H. Nash equilibrium seeking for multi-agent systems under DoS attacks and disturbances. *IEEE Transactions on Industrial Informatics*, 2024, **20**(4): 5395–5405
- 146 Wang J M, Zhang J F, He X K. Differentially private distributed algorithms for stochastic aggregative games. *Automatica*, 2022, **142**: Article No. 110440
- 147 Wang J M, Zhang J F. Differentially private distributed stochastic optimization with time-varying sample sizes. *IEEE Transactions on Automatic Control*, 2024, **69**(9): 6341–6348
- 148 Huang S J, Lei J L, Hong Y G. A linearly convergent distributed Nash equilibrium seeking algorithm for aggregative games. *IEEE Transactions on Automatic Control*, 2023, **68**(3): 1753–1759
- 149 Wang Jian-Rui, Huang Jia-Hao, Tang Yang. Swarm intelligence capture-the-flag game with imperfect information based on deep reinforcement learning. *Scientia Sinica Technologica*, 2023, **53**(3): 405–416  
(王健瑞, 黄家豪, 唐漾. 基于深度强化学习的 imperfect 信息群智夺旗博弈. 中国科学: 技术科学, 2023, **53**(3): 405–416)
- 150 Erez L, Lancewicki T, Sherman U, Koren T, Mansour Y. Regret minimization and convergence to equilibria in general-sum Markov games. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR, 2023. 9343–9373
- 151 Jin C, Liu Q H, Wang Y H, Yu T C. V-learning—A simple, efficient, decentralized algorithm for multiagent RL. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR, 2022.
- 152 Zhang B, Li L J, Xu Z W, Li D P, Fan G L. Inducing stackelberg equilibrium through spatio-temporal sequential decision-making in multi-agent reinforcement learning. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2023. Article No. 40
- 153 Mao W C, Yang L, Zhang K Q, Basar T. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 15007–15049
- 154 Ding D S, Wei C Y, Zhang K Q, Jovanovic M. Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 5166–5220
- 155 Cui Q W, Zhang K Q, Du S S. Breaking the curse of multiagents in a large state space: RL in Markov games with independent linear function approximation. In: Proceedings of the 36th Annual Conference on Learning Theory. Bangalore, India: PMLR, 2023. 2651–2652
- 156 Bakhtin A, Wu D J, Lerer A, Gray J, Jacob A P, Farina G, et al. Mastering the game of no-press diplomacy via human-regularized reinforcement learning and planning. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR, 2023.
- 157 Tennant E, Hailes S, Musolesi M. Modeling moral choices in social dilemmas with multi-agent reinforcement learning. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2023. Article No. 36
- 158 Yang D K, Yang K, Wang Y Z, Liu J, Xu Z, Yin R B, et al. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 1093



**李艺春** 华东理工大学信息科学与工程学院博士后。2023 年获得山东大学数学学院博士学位。主要研究方向为多智能体博弈决策, 多智能体强化学习与最优控制。

E-mail: [yichunli953@gmail.com](mailto:yichunli953@gmail.com)

(**LI Yi-Chun** Postdoctor at the School of Information Science and Engineering, East China University of Science and Technology. She received her Ph.D. degree from the School of Mathematics, Shandong University in 2023. Her research interest covers game and decision-making of multi-agent, multi-agent reinforcement learning and optimal control.)



**刘泽娇** 华东理工大学数学学院博士研究生。主要研究方向为多智能体强化学习。

E-mail: [liuzejiao@mail.ecust.edu.cn](mailto:liuzejiao@mail.ecust.edu.cn)

(**LIU Ze-Jiao** Ph.D. candidate at the School of Mathematics, East China University of Science and Technology. Her main research interest is multi-agent reinforcement learning.)



**洪艺天** 华东理工大学信息科学与工程学院博士研究生。主要研究方向为多智能体强化学习及其应用。

E-mail: [y20200105@mail.ecust.edu.cn](mailto:y20200105@mail.ecust.edu.cn)

(**HONG Yi-Tian** Ph.D. candidate at the School of Information Science and Engineering, East China

University of Science and Technology. His research interest covers multi-agent reinforcement learning and its application.)



**王继超** 华东理工大学信息科学与工程学院硕士研究生. 主要研究方向为多智能体强化学习.

E-mail: [jichaowang@mail.ecust.edu.cn](mailto:jichaowang@mail.ecust.edu.cn)  
(**WANG Ji-Chao** Master student at the School of Information Science and Engineering, East China

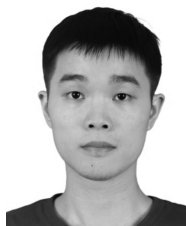
University of Science and Technology. His main research interest is multi-agent reinforcement learning.)



**王健瑞** 华东理工大学信息科学与工程学院博士研究生. 主要研究方向为多智能体强化学习, 博弈论.

E-mail: [jianruiwang@mail.ecust.edu.cn](mailto:jianruiwang@mail.ecust.edu.cn)  
(**WANG Jian-Rui** Ph.D. candidate at the School of Information Science and Engineering, East China

University of Science and Technology. His research interest covers multi-agent reinforcement learning and game theory.)



**李毅** 华东理工大学信息科学与工程学院博士研究生. 主要研究方向为多智能体强化学习.

E-mail: [Y13220018@mail.ecust.edu.cn](mailto:Y13220018@mail.ecust.edu.cn)  
(**LI Yi** Ph.D. candidate at the School of Information Science and Engineering, East China University

of Science and Technology. His main research interest is multi-agent reinforcement learning.)



**唐漾** 博士, 华东理工大学信息科学与工程学院教授. 主要研究方向为智能无人系统, 工业智能, 具身智能, 机器视觉, 强化学习. 本文通信作者.

E-mail: [yangtang@ecust.edu.cn](mailto:yangtang@ecust.edu.cn)  
(**TANG Yang** Ph.D., professor at the School of Information Science

and Engineering, East China University of Science and Technology. His research interest covers intelligent unmanned systems, industrial intelligence, embodied artificial intelligence, computer vision, and reinforcement learning. Corresponding author of this paper.)