中国科学: 信息科学

开放环境下的协作多智能体强化学习进展综述

A Survey of Progress on Cooperative Multi-agent Reinforcement Learning in Open Environment

期刊：	中国科学: 信息科学
稿件ID：	SSI-2023-0335
稿件栏目：	评述
作者提交日期：	2023-11-01
参与作者列表：	袁雷, 张子谦, 李立和, 管聪, 俞扬
关键词：	
英文关键词：	
学科领域：	人工智能理论

中国科学：信息科学

SCIENTIA SINICA Informationis

评述

开放环境下的协作多智能体强化学习进展综述

袁雷，张子谦，李立和，管聪，俞扬*

南京大学计算机软件新技术国家重点实验室，南京 210023

* 通信作者. E-mail: yuy@nju.edu.cn

国家自然科学基金创新研究群体项目（61876077）资助

摘要 多智能体强化学习（Multi-agent Reinforcement Learning, MARL）近年来获得广泛关注并在不同领域取得进展。其中，协作多智能体强化学习专注于训练智能体团队以协同完成单智能体难以应对的任务目标，在路径规划、无人驾驶、主动电压控制和动态算法配置等场景展现出巨大的应用潜力。如何提升系统协作效能是协作多智能体强化学习领域研究重点之一，以往的研究工作主要在简单、静态和封闭的环境设定中展开。随着人工智能技术落地的驱使，目前在多智能体协作领域也有部分研究开始对开放环境下的多智能体协作展开研究，这些工作从多个方面对智能体所处环境中要素可能发生改变这一情况进行探索与研究，并取得一定进展。但是当前主流工作仍然缺乏对该方向的综述。本文从强化学习概念着手，针对多智能体系统、协作多智能体强化学习、典型方法与测试环境进行介绍，对封闭到开放环境下的协作多智能体强化学习研究工作进行总结，提炼出多类研究方向并对典型工作进行介绍。最后，本文对当前研究的优势与不足进行了总结，对未来开放环境下协作多智能体强化学习的发展方向与待研究问题进行展望，以吸引更多研究人士参与这个新兴方向的研究与交流。

关键词 强化学习，多智能体系统，多智能体协作，开放环境机器学习，开放环境多智能体协作

1 引言

作为机器学习的一个分支，强化学习（Reinforcement Learning, RL）[1] 是一种解决序贯决策问题的有效方法，相较于监督学习与无监督学习，其显著的特点在于从交互中进行学习。在强化学习范式中，智能体通过与环境进行交互，根据所获得的奖赏或惩罚不断优化其策略。由于其学习方式与人类获取知识的方式类似，强化学习被视为实现通用人工智能（Artificial General Intelligence, AGI）[2] 的重要途径之一。早期的强化学习工作依赖于手工特征输入到线性模型进行估值和拟合，在复杂场景中表现不佳。在过去的十年中，得益于深度学习 [3] 的蓬勃发展，深度强化学习在各行各

引用格式: 袁雷, 张子谦, 李立和, 等. 开放环境下的协作多智能体强化学习进展综述. 中国科学: 信息科学, 在审文章
Yuan L, Zhang Z, Li L, et al. A Survey of Progress on Cooperative Multi-agent Reinforcement Learning in Open Environment (in Chinese). Sci Sin Inform, for review

业取得瞩目成就。例如，深度 Q 网络（Deep Q-Network, DQN）[4] 在雅达利（Atari）视频游戏上超越了人类职业玩家；AlphaGo [5] 在围棋人机大战中击败围棋世界冠军李世石；AlphaStar [6] 在不完全信息即时战略游戏星际争霸 II 中战胜了人类顶尖职业选手；OpenAI Five [7] 在多人实时在线竞技游戏 Dota2 中表现出色；而 AI-Suphx [8] 在多人非完全信息博弈麻将中也取得了重要的成果。此外，强化学习的应用范围也逐渐从游戏领域扩展到现实生活的各个领域，包括工业制造、机器人控制、物流管理、国防军事、智慧交通、智能医疗等，极大地推动了人工智能技术的发展 [9,10]，例如，最近得到广泛关注的 ChatGPT [11] 也使用了强化学习技术进行优化。近年来，在人工智能应用于科学研究（AI4Science）[12] 的发展趋势下，强化学习在许多基础科学领域也大放异彩，如 DeepMind 利用深度强化学习实现了核聚变的控制 [13]，发现矩阵乘法的强化学习方法 AlphaTensor [14] 等。

与此同时，现实中的很多问题往往是大规模、复杂、实时并带有不确定性的，将此类问题建模为单智能体问题效率低下且与现实条件不符，而将其建模为多智能体系统（Multi-agent System, MAS）[15] 问题则往往更加适配。进一步，由于很多复杂问题难以由某个个体独立完成，需要多智能体协同，如自动驾驶汽车、智能仓储系统和传感器网络等。协作多智能体强化学习（Multi-Agent Reinforcement Learning, MARL）[16∼18] 为这些问题的建模和解决提供了强有力的支撑。在这类问题中，智能体团队通过与环境交互学习一个联合的协作策略以解决任务。相较于传统算法，多智能体强化学习的优势在于能够应对环境的不确定性，并在不需要过多领域知识的条件下学习解决未知任务。近年来，深度学习与多智能体强化学习的结合产出了丰硕的成果 [19]，许多算法被提出并用于解决复杂任务。然而，多智能体强化学习区别于单智能体问题的特殊性也带来了新的挑战。一方面，真实多智能体系统所处环境往往是部分可观测的，单个智能体无法从其局部观测中获得环境的全局信息，这意味着独立学习的智能体难以做出最优决策 [20]。另一方面，由于其他智能体同时进行学习，策略相应地会发生变化，从单个智能体角度来看，其处于一个非稳态的环境中，收敛性无法得到保证 [21]。此外，协作型多智能体系统往往只能得到共享奖赏，如何将其分配从而为每个智能体提供准确的反馈，进而使其高效地学习协作，最终让系统性能得到最大提升也是目前存在的核心挑战之一 [22]。最后，随着多智能体系统的扩充，智能体的数量不断增加，解决强化学习问题所面临的搜索空间将呈指数增长，这使得在策略空间中搜索和学习变得极为困难。因此，如何组织高效的策略学习，也是目前的一大难点 [23, 24]。

针对前述挑战，目前的大量工作从多方面展开研究，并在许多任务场景中取得令人惊喜的成果。在路径规划 [25]、主动电压控制 [26] 和动态算法配置 [27] 等任务中，协作型多智能体强化学习都展现出比传统方法更优秀的性能。研究者设计出许多方法以促进智能体之间的协作，包括基于策略梯度的方法，如 MADDPG [28] 和 MAPPO [29]；基于值函数的方法，如 VDN [30] 和 QMIX [31]；或包括借助 Transformer 的强大表达能力提升协作能力的 MAT [32] 在内的其他方法。这些方法已经在许多任务（如 SMAC [33]、Hanabi 和 GRF [29]）中表现出优秀的协作能力。除了上述方法和相应的变体外，研究者亦从其他角度对协作型 MARL 进行了深入探索与研究，包括通过高效通信 [20] 缓解分布式策略执行设定下的部分可观测性、对策略进行离线部署 [34]、MARL 中的世界模型学习 [35] 和训练范式研究 [36] 等。

传统的机器学习研究通常在经典封闭环境情景这一假设下进行，其中学习过程的重要因素保持不变。如今越来越多的实际任务，尤其是涉及到开放环境情景的任务，其中的重要学习因素可能会发生变化。显然，对于机器学习来说，从经典环境转向开放环境是一个巨大的挑战。针对数据驱动

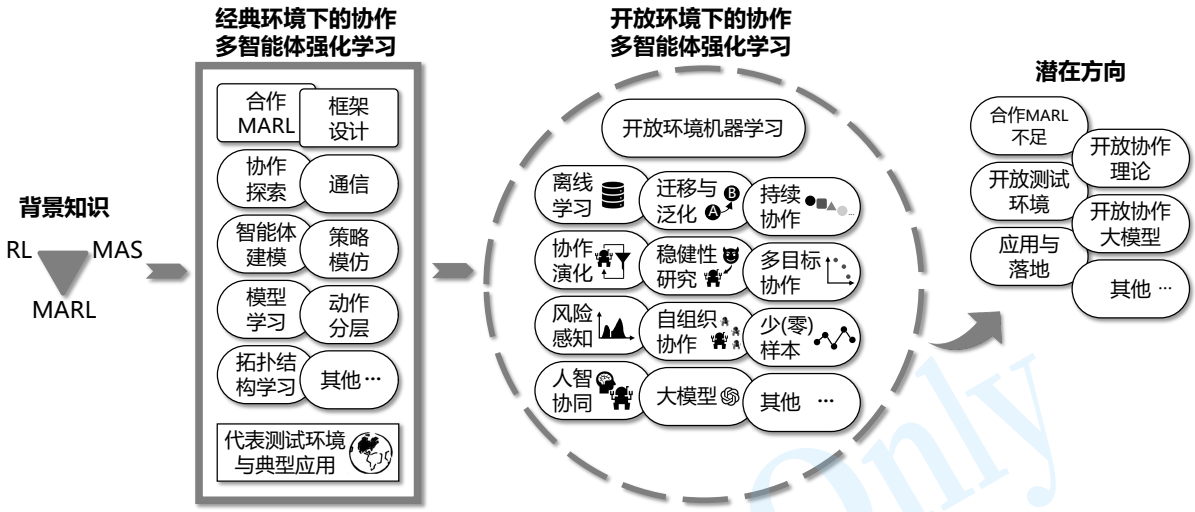

图 1 本文整体架构。

的学习任务，开放环境下数据会随着时间在线地积累，例如数据流这一形式，使得模型的学习更具挑战性。开放环境下的机器学习 [37,38] 在许多场景具有应用前景，逐渐引起广泛关注。在机器学习领域，目前的研究工作主要包括类别变化、特征增减、数据分布变化、学习目标变化等。与之对应，强化学习领域的部分工作也开始着眼于开放环境下的任务，主要工作包括可信强化学习 [39]、环境生成与策略学习 [40]、持续强化学习 [41]、强化学习泛化能力 [42]、元强化学习 [43] 与模拟器到真实环境的策略迁移 [44] 等。

相较于单智能体强化学习，多智能体场景更复杂也更有挑战性，目前对开放环境下的协作多智能体研究较少，有部分工作开始关注多智能体的稳健性研究 [45]，并从不同角度进行问题描述与算法设计 [46~49]。此外，针对开放团队的多智能体强化学习问题，一些工作提出了包括即时团队协作（Ad-Hoc Teamwork, AHT） [50]、零样本协作（Zero-Shot Coordination, ZSC） [51] 和少样本团队协作（Few-Shot Teamwork, FST） [52] 在内的设定，以应对这一挑战。虽然上述工作在一些任务场景取得了成效，然而其与大部分真实应用场景仍然不相匹配并且实际效果存在较大提升空间。目前针对多智能体强化学习已经存在部分综述性工作，比如对多智能体系统 [15]、多智能体强化学习 [16,53~58]、多智能体建模 [59]、多智能体中的非稳态处理 [21]、多智能体中的迁移学习 [60]、协作多智能体强化学习 [17,61,62]、多智能体模型学习 [35]、因果多智能体强化学习 [63]、多智能体通信 [20] 等。此外，部分工作也对开放机器学习进行了综合分析 [37,38,64]。虽然上述工作从多智能体强化学习亦或是开放环境机器学习方面都进行了一些回顾和归纳，但是针对开放环境下的多智能体强化学习，尤其开放环境下的协作多智能体强化学习，目前没有工作进行系统性综述。考虑到协作多智能体强化学习在解决真实环境中复杂协作问题的潜力与价值，本文拟对最近的工作进展进行描述。本文后续的安排如图 1 所示，我们首先介绍本文涉及到的背景知识，包括强化学习的基本知识、多智能体系统到多智能体强化学习常见知识与背景；之后，我们对经典环境下的协作多智能体强化学习进行介绍，从具体定义到当前主流研究内容，以及常见测试环境与应用案例；接下来，我们对开放环境下的协作多智能体强化学习进行介绍，具体而言，包括封闭机器学习、强化学习到协作多智能体常见研究方向与内容；最后我们总结本文的主要内容，并对开放环境下的协作多智能体

协作进行展望，以期为该方向研究思路与方法的拓展提供抛砖引玉的作用，引起更多研究者对开放环境下的协作多智能体强化学习的关注并展开研究。

2 背景知识

2.1 强化学习

强化学习 [1] 旨在指导智能体依据当前的状态学习执行恰当的动作，即学习到一个观测状态到动作的映射，通过决策以最大化环境反馈的累积数值奖赏（reward）。环境会依据当前状态和智能体执行的动作反馈其相应的奖赏信息。智能体无法得知要执行的最优动作，而必须通过尝试从而发现那些能产生最大累积奖赏的动作。在标准 RL 场景中，智能体通过观测状态和执行动作与环境交互。在交互的每个时间步，智能体都会接收对环境当前状态的观测，然后依据该观测选择动作作为输出。上述动作的执行会改变环境的状态，并发送给智能体环境反馈的奖赏信号。智能体的目的是执行能最大化累积数值奖赏的动作序列。

2.1.1 问题设定

强化学习 [1] 是机器学习的一个分支，相较于机器学习中经典的监督学习和无监督学习，其最大特点是在交互中学习，即智能体在与环境的交互中根据获得的奖赏或惩罚不断优化策略进而适应新的环境。强化学习主要由四个部分构成：智能体（Agent）、状态（State）、动作（Action）和奖赏（Reward）（图2）。强化学习的目标是最大化累积奖赏，智能体需要通过反复探索试错并学习，找到对应的最优策略，该过程一般可以建模为马尔可夫决策过程（Markov Decision Process, MDP）。

定义 1 (马尔可夫决策过程) 马尔可夫决策过程由五元组 $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$（无限过程）或五元组 $\langle \mathcal{S}, \mathcal{A}, P, R, T \rangle$（有限过程）组成，其中

- $\mathcal{S}$ 为所有状态的集合，
- $\mathcal{A}$ 为所有动作的集合，
- $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ 为状态转移概率函数: $P(s' \mid s, a) = \Pr[S_{t+1} = s' \mid S_t = s, A_t = a]$，
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 为奖赏函数: $R(s, a) = \mathbb{E}[r_t \mid S_t = s, A_t = a]$，
- $\gamma \in [0, 1]$ 是折扣因子，T 是最大轨迹长度。

智能体可以自由地选择在一个状态上执行哪个动作，或者按照一定的概率分布来选择。这里强调“自由选择”，即是智能体的“能动性”，可以自主地改变动作的选择。对于一种固定下来的动作选择，本文称之为策略。具体而言，在任一时间步 t，智能体从环境中观测到当前状态 s_t，然后执行动作 a_t。该动作使得智能体所处的环境依据转移函数 $s_{t+1} \sim P(\cdot|s_t, a_t)$ 转移至新的状态，并且接收环境反馈的奖赏信息 $r_t = R(s_t, a_t)$。不失一般性，下文主要讨论无限过程的 MDP，智能体的目标是：找到最优策略以最大化累积奖赏。在数学上，上述过程可以总结为让智能体找到一个马尔可夫的（输出仅取决于当前状态）且平稳的（函数形式与时间无关）策略函数 $\pi : \mathcal{S} \to \Delta(\mathcal{A})$，从而引导智能体执行恰当的序列决策，最大化累积奖赏，其中 $\Delta(\mathcal{A})$ 是动作空间上的概率分布，优化目标即：

$$\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0\right], \tag{1}$$

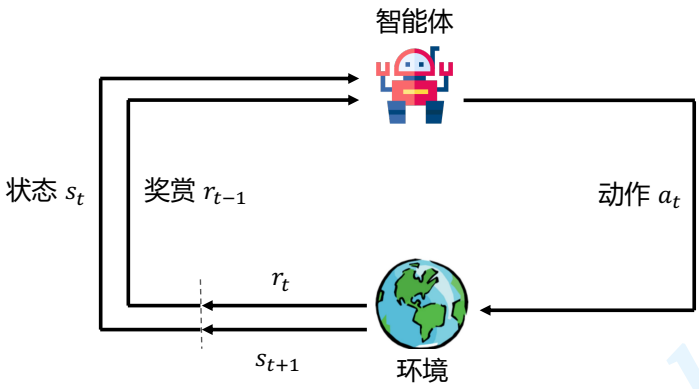

图 2 强化学习示意图。

其中的 $\mathbb{E}^{\pi}[\cdot]$ 是针对策略 $\pi(a_t|s_t)$ 和状态转移 $P(s_{t+1}|s_t,a_t)$ 下得到的序列 $\tau=(s_0,a_0,s_1,a_1,\ldots)$ 的分布计算所得的期望。

依据式 1 的目标函数，在策略 π 的指导下，可以定义状态-动作价值函数，即 Q-函数。它表示在策略 π 下，基于状态 s 执行动作 a 后的期望累积奖赏；此外，状态值函数 V 描述的是基于状态 s 的期望累积奖赏：

$$\begin{aligned} Q^{\pi}(s,a)&=\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)\mid s_0=s,a_0=a\right],\\ V^{\pi}(s)&=\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)\mid s_0=s\right]. \end{aligned} \tag{2}$$

显然，状态动作函数 Q-函数和状态价值函数 V-函数之间的关系可以表示为 $V^{\pi}(s)=\mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^{\pi}(s,a)\right]$ 和 $Q^{\pi}(s,a)=\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[R(s,a)+V^{\pi}(s')\right]$。基于两类值函数的定义，可以将强化学习在马尔可夫决策过程中的学习目标表示为寻找最优的策略 π_* 以最大化价值函数：$\pi_*=\arg\max_{\pi}V^{\pi}(s),\forall s$。

2.1.2 基于值函数的强化学习

对于状态空间和动作空间有限的马尔可夫决策过程，至少存在一个确定性平稳最优策略使累积奖赏最大化。而基于值函数的算法需要构建并估计值函数，该值函数随后用于指导动作的选择和执行，导出对应的策略。此类算法大多都不直接使用 V-函数，而需使用 Q-值函数，使得式 2 对应的 Q-值最大化。相应的，通过取贪心动作，由 Q-函数诱导出的最优策略可以表示为 $\pi^*(a|s)=\mathbb{1}\{a=\arg\max_a Q^*(s,a)\}$，其中 $\mathbb{1}\{\cdot\}$ 为指示函数。经典的 Q-学习算法通过时序/时间差分（temporal-difference, TD）学习代理函数 $\hat{Q}$-值 [65]，以近似最优函数 Q^*，更新过程为：

$$\hat{Q}(s_t,a_t)\leftarrow\hat{Q}(s_t,a_t)+\alpha\cdot\overbrace{\left(\underbrace{r_t+\gamma\max_{a_{t+1}\in\mathcal{A}}\hat{Q}(s_{t+1},a_{t+1})}_{\text{差分目标}}-\hat{Q}(s_t,a_t)\right)}^{\text{差分误差}}. \tag{3}$$

理论上，给定贝尔曼最优算子 $\mathcal{B}^*$ 后，Q-值函数的求解过程可以被定义为 [66]：

$$(\mathcal{B}^*Q)(s,a)=\sum_{s'}P(s'|s,a)\left[R(s,a)+\gamma\max_{a'\in\mathcal{A}}Q(s',a')\right], \tag{4}$$

为了建立最优的 Q-值函数，Q-学习算法利用贝尔曼方程的不动点迭代来求解具有（唯一）解的 Q-值函数 $Q^*(s,a)=(\mathcal{B}^*Q^*)(s,a)$。在实践中，当环境模型不可知，状态-动作对均为离散表示，且所有动作在所有状态下可重复采样的情况下，上述 Q-学习方法可以保证收敛到最优解。

由于许多现实问题的环境可能具有连续的、高维的状态-动作空间，导致上述假设通常不适用。此时，智能体需要对参数化的 Q-值函数 $Q(s,a|\theta)$ 进行学习，其中 θ 是函数实例化的参数。为了更新 Q-函数，智能体需要以元组 (s,a,r,s') 的形式收集其与环境交互时产生的样本序列。其中奖赏 r 和下一个时间步的状态 s' 遵循环境基于状态-动作对 (s,a) 的反馈。每次迭代时，当前的 Q-值函数 $Q(s,a|\theta)$ 可以进行如下更新 [67]：

$$\begin{aligned} y^Q&=r+\gamma\max_{a'\in\mathcal{A}}Q(s',a'|\theta),\\ \theta&\leftarrow\theta+\alpha\left(y^Q-Q(s,a|\theta)\right)\nabla_\theta Q(s,a|\theta),\end{aligned} \tag{5}$$

其中的 α 是更新速率。

式 5 中的 Q-学习方式可以直接使用神经网络 $Q(s,a|\theta)$ 拟合，以向最优 Q^*-值收敛，其中参数 θ 通过随机梯度下降（或其他优化方式）更新。然而，由于神经网络的泛化和推理能力有限，Q-网络在状态-动作空间的不同位置会产生不可预测的变化。因此，式 4 中的贝尔曼算子的收缩映射性质不足以保证收敛性。大量实验证明，这些误差会随着在线更新规则的传播而传导，从而导致收敛速度缓慢甚至不稳定。使用函数逼近 Q-值的另一不利影响是，由于 max 算子的作用，Q-值往往会被过高估计。因此，由于不稳定性和过高估计的风险，必须特别注意保证适当的学习速率。Mnih 等人提出的深度 Q 网络（Deep Q Network，DQN） [68] 学习算法使用了两个重要元素以缓解上述不利影响：目标 Q-网络和经验回放/重放。具体来说，DQN 实例化一个参数为 θ^- 的目标 Q-网络，用以计算差分目标；同时利用经验回放池 $\mathcal{D}$ 存储采样收集的样本序列，不仅保证了样本利用率，同时在训练时通过独立同分布的采样以缓解采样数据的相关性。DQN 的优化目标由此可以表示为：

$$\min_\theta\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}[(r+\gamma\max_{a'\in\mathcal{A}}Q(s',a'|\theta^-)-Q(s,a|\theta))^2], \tag{6}$$

其中目标 Q-网络 $Q(s,a|\theta^-)$ 会定期地更新以同步与 Q-网络的参数。通过上述方式以及两个重要元素可以实现对 Q-网络的稳定训练。

此外，近期有许多工作对 DQN 进行了额外的改进：在 [69] 中，利用神经网络结构解耦了值函数和优势函数（定义为 $A(s,a)=Q(s,a)-V(s)$），从而提高了学习性能。在 [70] 中，一种特定的更新方案（公式6 的变体）可以减少对 Q-值的高估，同时提高学习的稳定性。并行学习 [71] 或使用无监督辅助任务 [72] 也有助于更快、更稳健的学习。在 [73] 中，可微记忆模块可以通过在蒙特卡罗值估计和异策略估计之间进行插值，从而快速集成最近的经验。

2.1.3 基于策略梯度的强化学习

基于策略的方法（Policy-based Methods） [74] 的基本思想是直接学习参数化的最优策略 π_θ。与基于值函数的方法相比，基于策略的算法具有以下特性：首先，这类方法可以灵活地应用于连续动

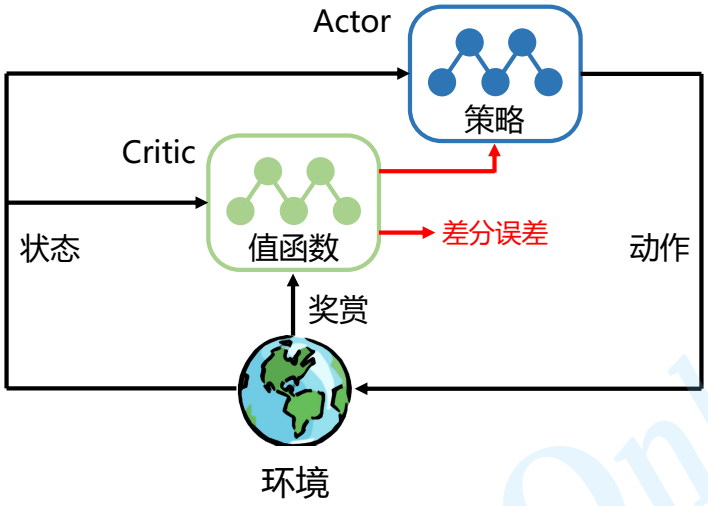

图 3 Actor-Critic 方法体系结构。

作空间中；另外，此类方法可以直接获得随机策略 $\pi_\theta(\cdot|s)$。当用神经网络参数化策略时，典型方法是朝着提高累积奖赏的方向调整参数：$\theta \leftarrow \theta + \alpha \nabla_\theta V^{\pi_\theta}(s)$。然而，梯度也将受制于策略变化对状态分布的未知影响。早在 [1] 一文中，研究者基于策略梯度（Policy Gradient）定理推导出了一个不涉及状态分布的解：

$$\nabla_\theta V^{\pi_\theta}(s) = \mathbb{E}_{s\sim\mu^{\pi_\theta}(\cdot), a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) \cdot Q^{\pi_\theta}(s,a)\right], \tag{7}$$

其中的 μ^{π_θ} 是在策略 π_θ 下状态的占用度量 [75]，$\nabla_\theta \log \pi_\theta(a|s)$ 是策略的更新得分评估。当策略是确定性，且动作空间是连续的，我们便可进一步得到确定性策略梯度（Deterministic Policy Gradient，DPG）定理 [76]：

$$\nabla_\theta V^{\pi_\theta}(s) = \mathbb{E}_{s\sim\mu^{\pi_\theta}(\cdot)} \left[\nabla_\theta \pi_\theta(a|s) \cdot \nabla_a Q^{\pi_\theta}(s,a)|_{a=\pi_\theta(s)}\right]. \tag{8}$$

关于策略梯度最经典的应用是 REINFORCE 算法 [77]，其使用累积奖赏 $R_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$ 作为对 $Q^{\pi_\theta}(s_t, a_t)$ 的估计来对策略参数进行更新。

2.1.4 基于行动者-评论者的强化学习方法

传统的行动者-评论者（Actor-Critic）方法由两个部分组成：Actor，用以调整策略 π_θ 的参数 θ；以及 Critic，用来调整状态-动作值函数 $Q_w^{\pi_\theta}$ 的参数 w。基于前述两类方法，相应地也可以得到关于 Actor 和 Critic 的更新方式分别为：

$$\begin{aligned} \theta &\leftarrow \theta + \alpha_\theta Q_w(s,a) \nabla_\theta \log \pi_\theta(s,a), \\ w &\leftarrow w + \alpha_w (r + \gamma Q_w(s',a') - Q_w(s,a)) \nabla_w Q_w(s,a). \end{aligned} \tag{9}$$

Actor-Critic 方法将策略梯度方法和值函数近似的方法两者结合，Actor 基于概率选择动作，Critic 基于 Actor 选择的动作以及当前状态评判该动作的得分，然后 Actor 根据 Critic 的评分修改其

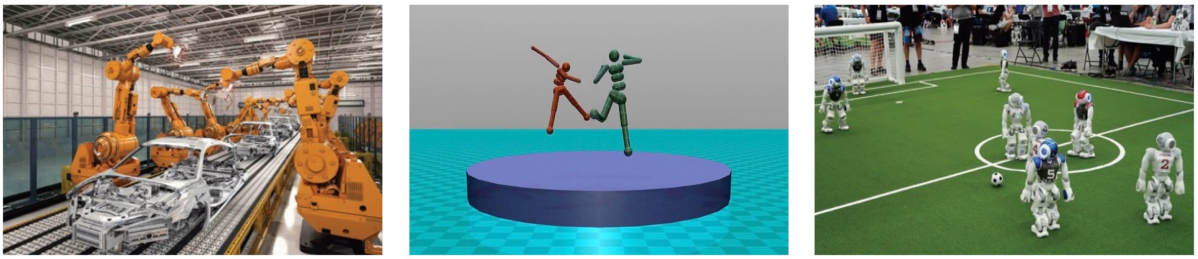

图 4 多智能体系统的三种常见设定。

选择动作的概率。其优势在于此类方法可以进行单步更新，在连续动作空间中得到得到低方差的解 [78, 79]，而代价则是学习开始时，由于 Critic 的估计不够准确，算法有较大的波动。图 3 描述了 Actor-Critic 方法的体系结构，更多关于单智能体强化学习的方法可以参见综述 [9]。

2.2 多智能体强化学习

2.2.1 多智能体系统

多智能体系统由分布式人工智能（Distributed Artificial Intelligence，DAI）演变而来 [80]，其研究目的是解决大规模、复杂、实时和有不确定性的现实问题，此类问题通过单智能体建模往往会效率低下并且与现实条件相悖。多智能体系统具有自主性、分布性、协调性等特点，并具有自组织能力、推理能力和学习能力。自 20 世纪 70 年代被提出以来，经过多年的发展，目前多智能体系统的研究既包括构建单个智能体的技术，如建模、推理、学习及规划等，也包括使多个智能体协调运行的技术，例如交互通信、协调、合作、协商、调度、冲突消解等。多智能体系统在智能机器人、交通控制、分布式决策、软件开发、游戏等领域都得到迅速而广泛的应用，目前已经成为一种对复杂系统进行分析与模拟的工具 [15]。

依据任务特性的不同，一般可以将多智能体系统划分为完全协作（Fully Cooperative Setting）、完全竞争（Fully Competitive Setting）和混合关系（Mixed setting）这三种设定。其中完全协作关系指系统中的智能体拥有一个相同的目标，智能体相互协作完成特定任务，例如机器系统装配；完全竞争关系下，智能体间的目标冲突，一方的收益导致另一方的损失，例如相扑对抗中的双方；混合关系指智能体间可以划分为多个群组，组内智能体相互合作，而组间是相互竞争的关系，例如足球比赛中，队友间是协作关系而队与队之间是竞争关系。

多智能体系统中，不同设定下的策略优化目标有差别，其所学出的策略也会有所不同。在协作的设定下，每个智能体的决策要考虑到队友的策略，与队友做到尽量好的配合，避免产生干扰。而在竞争的设定下，智能体要考虑到对手的策略，最小化对手同时最大化自己的回报，并以此为目标相应优化自身策略。

2.2.2 博弈论简介

博弈论（game theory） [81, 82] 作为一种分析理性智能体之间策略交互的数学理论与方法，最早被广泛应用在经济学领域，并逐渐覆盖社会学、政治、心理学以及计算机科学等多个学科。为此，

我们可以将多智能体系统中个体之间的协作与竞争建模为博弈问题，分析智能体之间的策略选择和决策制定过程。博弈论中通常将问题建模为正则博弈，可定义为如下形式:

定义 2 (正则式博弈（Normal-form Games）) 一个（有限，n-player）正则式博弈由三元组 $(\mathcal{N}, \mathcal{A}, \boldsymbol{Q})$ 组成，其中

- $\mathcal{N}$ 是一个有限的智能体的集合，其大小为 n，每个智能体用 $i \in \{1, \cdots, n\}$ 索引表示;
- $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ 表示动作组合空间，其中每个 $\mathcal{A}_i$ 表示智能体 i 的可选动作集合，而 $\boldsymbol{a} = (a_1, a_2, \cdots, a_n) \in \mathcal{A}$ 被称为一个动作组合（action profile）。
- $\boldsymbol{Q} = (Q_1, \cdots, Q_n)$, 其中每个分量 $Q_i : \mathcal{A} \mapsto \mathbb{R}$ 是智能体 i 的效益函数 (utility or payoff function, 也称收益函数)。

虽然博弈论中常用符号 u 来标识效益函数，但是其与单状态，轨迹长度为 1 情况下的 MDP 中的动作-价值函数意义相同，故我们以符号 Q 进行替代以保证符号的统一。当博弈问题中仅有两个参与者，且这两个智能体处于完全竞争场景，一般可以建模为二人零和博弈过程:

定义 3 (二人零和博弈) 当收益函数满足 $\forall \boldsymbol{a}, Q_1(\boldsymbol{a}) + Q_2(\boldsymbol{a}) \equiv 0$ 时，称为二人零和博弈。

在正则博弈中, 如果智能体采取某个单一动作执行, 此时我们称智能体执行了一个纯策略 (pure strategy，也即确定性策略)。当每个智能体都执行单一动作时，得到的动作组合 $\boldsymbol{a}$ 通常被称为纯策略组合（pure strategy profile)。更一般地，智能体的策略可定义为:

定义 4 (策略与策略组合) 在正则博弈 $(\mathcal{N}, \mathcal{A}, \boldsymbol{Q})$ 中，智能体 i 的（混合）策略是在其可行动作空间 $\mathcal{A}_i$ 上定义的一个概率分布，$\pi_i : \mathcal{A}_i \mapsto [0, 1]$。所有智能体的策略的笛卡尔积 $\boldsymbol{\pi} = \prod_{i=1}^{n} \pi_i$ 是一个 (混合) 策略组合。

同样的，我们以强化学习所使用的符号 π_i 代替博弈论中的常用符号 σ_i，来表示在可行动作空间上的概率分布。此外，在博弈论中，通常用 $-i$ 表示除了智能体 i 以外的其他人的集合。由此，在混合策略组合 $\boldsymbol{\pi} = (\pi_i, \boldsymbol{\pi}_{-i})$ 下的智能体 i 的期望收益为

$$Q_i(\pi_i, \boldsymbol{\pi}_{-i}) = \sum_{\boldsymbol{a} \in \mathcal{A}} Q_i(\boldsymbol{a}) \prod_{i=1}^{n} \pi_i(a_i).$$

从期望收益的形式可以看出，对智能体 i 而言，二人常和博弈与二人零和博弈的期望收益函数仅相差常数 c，因此，两种博弈在相同解概念下的策略不会有区别。对于理性的智能体而言，其目标往往是追求利益最大化，即最大化自己的期望收益。因此，我们需要先提出最优反应的定义。

定义 5 (最优反应（Best Response）) 如果智能体 i 的策略 $\pi_i^\star$ 对其他智能体 $-i$ 的策略组合 $\boldsymbol{\pi}_{-i}$，满足

$$\forall \pi_i', Q_i(\pi_i^\star, \boldsymbol{\pi}_{-i}) \geqslant Q_i(\pi_i', \boldsymbol{\pi}_{-i}),$$

则称策略 $\pi_i^\star$ 是针对 $\boldsymbol{\pi}_{-i}$ 的一个最优反应，对 $\boldsymbol{\pi}_{-i}$ 的最优反应的集合记为 $BR(\boldsymbol{\pi}_{-i})$。

最优反应的直观含义是，在给定其他智能体策略 $\boldsymbol{\pi}_{-i}$ 的情况下，使得自己收益最大的策略。因此，如果固定了其他智能体的策略，便可使用强化学习求解最优反应。基于最优反应，我们可以很直接地提出纳什均衡的定义:

定义 6 (纳什均衡（Nash Equilibrium）) 在有 n 个智能体的博弈中，如果一个策略组合 $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_n)$ 满足 $\forall i \in \{1, \cdots, n\}, \pi_i \in BR(\boldsymbol{\pi}_{-i})$，则 $\boldsymbol{\pi}$ 是一个纳什均衡。

根据前述定义，在纳什均衡中，某个智能体单方面的偏离不会提高自己的收益，因此每个理性的智能体都没有单方面偏离均衡的动机。除了纳什均衡以外，在博弈论中，还有很多其他的解概念，例如极大极小策略（maxmin strategy）以及极小极大策略（minmax strategy）。

定义 7 (Maxmin 策略) 智能体 i 的极大极小 (maxmin) 策略为 $\arg\max_{\pi_i}\min_{\boldsymbol{\pi}_{-i}} Q_i(\pi_i,\boldsymbol{\pi}_{-i})$，其对应的极大极小值（maxmin value）是 $\max_{\pi_i}\min_{\boldsymbol{\pi}_{-i}} u_i(\pi_i,\boldsymbol{\pi}_{-i})$。

Maxmin 策略最大化了智能体 i 在最坏情况下（最有恶意的对手）的收益，因此，如果智能体 i 采用 maxmin 策略时，可以保证其收益不低于 maxmin 值。当不对其他智能体进行任何假设，单纯试图最大化自己的收益时，maxmin 策略是一个非常合适的选择，当然，这也是最保守的一种策略。与 maxmin 策略的“对偶”策略是 minmax 策略。

定义 8 (Minmax 策略) 其他智能体 $-i$ 的 minmax 策略为 $\arg\min_{\boldsymbol{\pi}_{-i}}\max_{\pi_i} Q_i(\pi_i,\boldsymbol{\pi}_{-i})$，此时智能体 i 对应的极小极大值（minmax value）是 $\min_{\boldsymbol{\pi}_{-i}}\max_{\pi_i} Q_i(\pi_i,\boldsymbol{\pi}_{-i})$。

上述两种策略往往在常和博弈中被应用。然而在协作中，我们更关注系统中的智能体通过某些协调手段选择策略以实现共同受益。其中，相关均衡提供了合理的策略协调机制：

定义 9 (相关均衡) 一个联合策略 $\boldsymbol{\pi}$ 是相关均衡，当且仅当对 $\forall i\in\mathcal{N},\forall a^i\in\mathcal{A}^i$

$$\sum_{\boldsymbol{a}^{-i}}\boldsymbol{\pi}(a^{i,*},\boldsymbol{a}^{-i})[Q_i(a^{i,*},\boldsymbol{a}^{-i})-Q_i(a^i,\boldsymbol{a}^{-i})]\geqslant 0,$$

其中 $a^{i,*}$ 是 $\boldsymbol{a}^{-i}$ 的最优反应。

相关均衡阐述了这样一个事实，在假设两个智能体服从一个相关联的策略分布的前提下，每个智能体不能改变当前策略而获得更高的效用价值，需要说明的是，纳什均衡下，每个智能体做出的选择都是独立的，也即智能体的行为选择不相关联，它可以被视作相关均衡的一个特殊情况。

前述概念都假设智能体同时做出决策，但在某些情景下，可能存在决策的先后顺序。对此，我们将智能体定义为领导者和追随者，领导者先做出决策，追随者随后做出决策。因而，领导者在决策时会有先发优势 (First-Mover Advantage)，可以通过预测追随者对其决策的反应来决定能够给自身带来最大收益的最佳决策。

定义 10 (斯塔克尔伯格均衡) 假设在顺序执行动作的场景下，领导者 π_l 先做出策略动作，追随者 π_f 后做出策略动作。Q_l 与 Q_f 分别为领导者和追随者的效用函数，那么斯塔克尔伯格均衡 (π_l^*,π_f^*) 满足以下约束：

$$Q_l(\pi_l^*,\pi_f^*)\geqslant Q_l(\pi_l,\pi_f^*),$$

其中

$$\pi_f^*=\arg\max_{\pi_f} Q_f(\pi_l,\pi_f).$$

2.2.3 多智能体强化学习

多智能体学习 (Multi-Agent Learning，MAL) 将机器学习技术引入多智能体系统领域，研究如何设计算法来创建动态环境下的自适应智能体学习与优化。多智能体强化学习 (Multi-Agent Reinforcement Learning，MARL) （图 5 ）的目标是在共享环境中通过强化学习方法训练多个智能体，从而完成给定的任务 [19,56]。

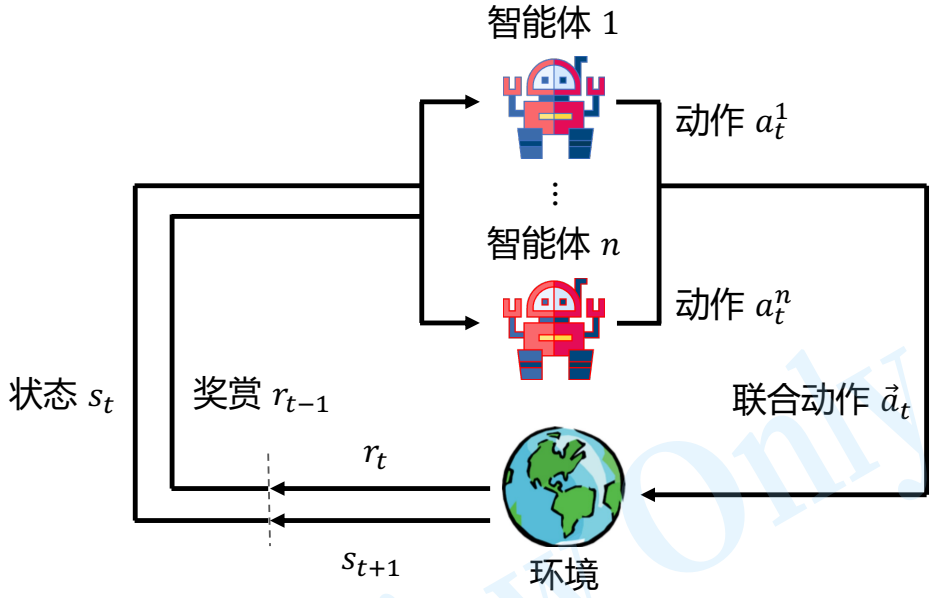

图 5 MARL 示意图。

与单智能体强化学习中将多步决策过程建模为 MDP 不同，多智能体强化学习一般被建模为随机博弈 (Stochastic Game)。

定义 11 (随机博弈) 一个随机博弈一般可以由一个六元组 $\langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, P, \{R^i\}_{i\in\mathcal{N}}, \gamma\rangle$ 构成，其中:

- $\mathcal{N} = \{1, 2, ..., n\}$ 代表系统中的智能体集合，当 $n = 1$ 时，问题退化成单智能体马尔可夫决策过程，$n \geqslant 2$ 是一般的随机博弈过程,
- $\mathcal{S}$ 是环境中所有智能体共享的状态空间,
- $\mathcal{A}^i$ 是智能体 i 的动作空间，定义联合动作空间 $\mathcal{A} := \mathcal{A}^1 \times \cdots \times \mathcal{A}^n$,
- $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ 是状态转移函数，具体而言，它刻画了在每个时间步，给定一个联合动作 $\boldsymbol{a} \in \mathcal{A}$，环境从 $s \in \mathcal{S}$ 转移到另外一个状态 $s' \in \mathcal{S}$ 的概率,
- $R^i : \mathcal{S} \times \mathcal{A} \times \mathcal{N} \to \mathbb{R}$ 是每个智能体的奖赏函数,
- $\gamma \in [0, 1]$ 是折扣因子。

在每一时间步，智能体 $i \in \mathcal{N}$ 处于状态 s，选择动作 $a^i \in \mathcal{A}^i$，构成联合动作 $\boldsymbol{a} = \langle a^1, \ldots, a^n\rangle$ 并执行，环境转移到下一个状态 $s' \sim P(\cdot \mid s, \boldsymbol{a})$，智能体 i 获得自身奖赏 $R^i(s, \boldsymbol{a})$。每个智能体通过优化自身的策略函数 $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}^i)$ 以最大化其期望累积奖赏，以状态价值函数的形式表达如下:

$$\max_{\pi_i} V^i_{\boldsymbol{\pi}}(s) := \mathbb{E}[\sum_{t\geqslant 0} \gamma^t R^i(s_t, \boldsymbol{a}_t) | a^i_t \sim \pi_i(\cdot \mid s_t),$$
$$\boldsymbol{a}^{-i}_t \sim \boldsymbol{\pi}_{-i}(\cdot \mid s_t), s_0 = s],$$

上述的符号 $-i$ 代表除了智能体 i 之外的其他智能体。与单智能体强化学习中智能体仅仅需要考虑自身对环境的影响不同，多智能体系统中智能体之间也会相互影响，他们共同决策并且同时更新策略。当系统中的其他智能体的策略固定时，智能体 i 可以最大化自己的收益函数，以寻找到相对于其他智

能体的策略的最优策略 π_i^*。在多智能体强化学习中，合理性（Rationality）与收敛性（Convergence）是学习算法最主要的评价指标。

定义 12 (合理性) 在对手使用一个恒定策略的情况下，当前智能体能够学习并收敛到一个相对于对手策略的最优策略。

合理性是一个相对基本的多智能体学习算法的性质，根据上述定义可知，该性质能够保证当一个智能体采用某种静态策略的时候，学习的智能体能收敛到均衡点。收敛性是另外一个基本指标，是算法具有收敛性能的保证。

定义 13 (收敛性) 在其他智能体也使用学习算法时，当前智能体能够学习并收敛到一个稳定的策略。通常情况下，收敛性针对系统中的所有智能体使用相同的学习算法。

由上述讨论内容可知，在多智能体强化学习的过程中，每个智能体希望增加自身的效用，此时的学习目标可以是最大化自身的 Q 函数。因此一个直观的学习方法是对每个智能体构建一个独立的 Q 函数，并按照 Q-学习算法进行更新：

$$Q_i(s_t, a_t^i) \leftarrow Q_i(s_t, a_t^i) + \alpha[r_{t+1} + \gamma \max_{a_{t+1}^i \in \mathcal{A}^i} Q_i(s_{t+1}, a_{t+1}^i) - Q_i(s_t, a_t^i)]. \tag{10}$$

与单智能体强化学习算法一样，每个智能体通过贪心策略选择当前执行动作：

$$\pi_i(a_t^i|s_t) = \mathbb{1}\{a_t^i = \arg\max_{a_t^i} Q_i(s_t, a_t^i)\} \tag{11}$$

上述情况中我们默认每个智能体都能够获取全局状态信息，从而进行决策。但是受限于现实条件，多智能体中的个体往往只能获得有限的局部观测。针对这一特性，我们一般将这样的决策过程建模为部分可观测随机博弈 (Partially-Observable Stochastic Games, POSG)：

定义 14 (部分可观测随机博弈) 定义为 $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, P, \{R^i\}_{i\in\mathcal{N}}, \gamma, \{\Omega^i\}_{i\in\mathcal{N}}, \mathcal{O}\rangle$，其前六项和随机博弈的定义 11 一致，除此之外，增加以下部分：

- Ω^i 是智能体 i 的观测空间，所有智能体的联合观测空间为 $\mathbf{\Omega} := \Omega^1 \times \cdots \times \Omega^n$,
- $\mathcal{O} : \mathcal{S} \to \Delta(\mathbf{\Omega})$ 代表了观测函数，$\mathcal{O}(\boldsymbol{o}|s)$，指在给定状态 s 后关于联合观测 $\boldsymbol{o}$ 的概率函数。

在部分可观测随机博弈中，每个智能体通过更新自己的策略 $\pi_i \in \Pi : \Omega^i \to \Delta(\mathcal{A}^i)$ 以最大化收益。虽然部分可观测随机博弈在真实场景中应用比较广泛，然而理论结果证明其求解难度是 NEXP-hard 的 [83]。所幸的是，近几年提出的包括集中式的多智能体强化学习训练框架、利用循环神经网络编码历史信息 [84] 和智能体通信 [20] 等技术可以从不同角度缓解该问题。

特别地，在协作型任务中，可以进一步将部分可观测随机博弈建模为分布式局部可观测马尔可夫决策过程（Decentralized Partially Observable MDP, Dec-POMDP）[83]，两者的主要区别在于后者中每个智能体的奖赏函数相同。除了部分可观测这一困难外，相较于单智能体强化学习，多智能体强化学习还面临着更多独有的问题。多智能体强化学习中，状态空间变大，动作空间随着智能体数量的增长出现“维度爆炸”（Curse of Dimensionality）问题 [56]。此外，包括可扩展性、信度分配、异质性、协同探索等问题的存在长期以来也在阻碍着多智能体强化学习的发展 [55]。

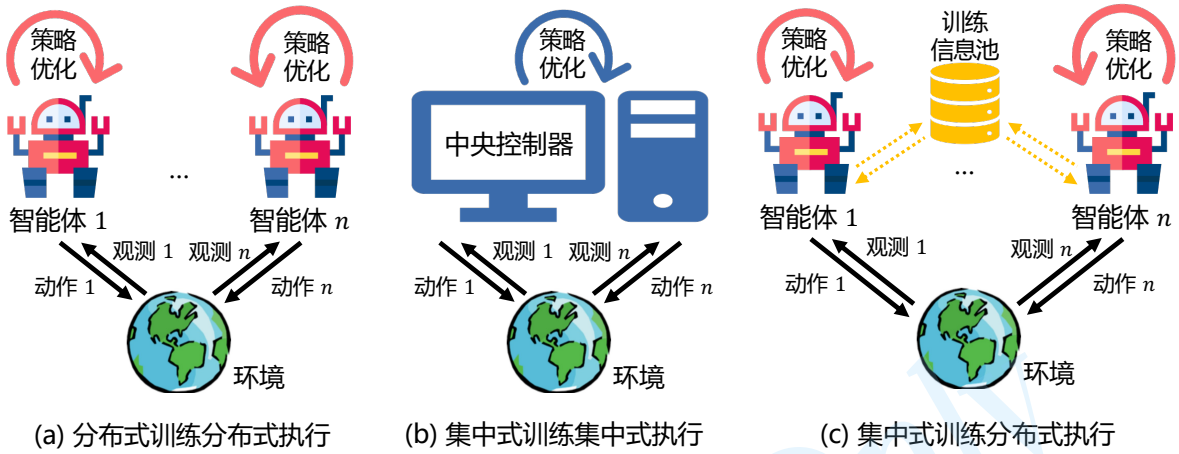

图 6 三种不同的训练范式。

2.2.4 多智能体强化学习训练范式

在多智能体强化学习中，训练指通过获取的经验（状态、动作、奖赏等）对策略进行优化的过程，执行指的是智能体根据自身或联合策略执行动作与环境交互。一般而言，根据智能体更新自身策略时是否需要其他智能体的信息，多智能体强化学习的训练过程可以划分为集中式训练（Centralized Training）与分布式训练（Decentralized Training）；与此相对应的是，根据在执行阶段是否需要外部信息，分为集中式执行（Centralized Execution）与分布式执行（Decentralized Execution）。结合训练阶段与执行阶段，多智能体强化学习可以分为三种范式，分布式训练分布式执行（Decentralized Training Decentralized Execution, DTDE), 集中式训练集中式执行（Centralized Training Centralized Execution, CTCE），集中式训练分布式执行（Centralized Training Decentralized Execution, CTDE），各结构如图 6 所示。

定义 15 (分布式训练分布式执行) 在 DTDE 框架下，每个智能体仅利用自己的局部信息独立地进行策略更新和策略实现，不涉及到信息的交换，

$$\pi_i : \Omega^i \to \Delta(\mathcal{A}^i).$$

IQL [85] 是基于 DTDE 的一个典型算法，由于此类算法仅需要考虑某个智能体本身，具有较强的可扩展性；同时，由于智能体不考虑其他智能体的信息或状态，智能体往往处于非稳态过程（见定义 18），无法学到最优的策略。为了解决分布式学习带来的效率低下问题，乐观优化与滞后优化通过对目前较好的 Q 赋予更大的更新权重，在一定程度上能缓解以上问题。与此同时，在通过 DQN 进行策略更新时，由于其他智能体的策略也进行了更新，直接使用经验回放池的数据会加剧非稳态，对此，可以通过重要性采样缓解此问题：

$$\mathcal{L}(\theta_i) = \frac{\boldsymbol{\pi}_{-i}^{t_c}(\boldsymbol{a}^{-i}|\boldsymbol{o}^{-i})}{\boldsymbol{\pi}_{-i}^{t_i}(\boldsymbol{a}^{-i}|\boldsymbol{o}^{-i})}[(y_i^{\mathrm{Q}} - Q_i(o^i, a^i; \theta_i))^2],$$

上式中的 θ_i 为智能体 i 的 Q 网络参数，t_c 为当前时刻，t_i 是样本的采集时间，y_i^Q 是时序差分目标，计算与单智能体部分所介绍类似。与此同时，使用循环神经网络等技术也可以在一定程度上缓解上

述问题。相较于分布式训练，集中式训练框架允许智能体在训练过程中进行信息交换，甚至可以拿到其他智能体的所有信息。

定义 16 (集中式训练集中式执行) 在 CTCE 框架下，智能体学习一个集中式的联合策略，

$$\boldsymbol{\pi}: \boldsymbol{\Omega} \to \Delta(\boldsymbol{\mathcal{A}}).$$

在 CTCE 框架下，我们可以使用任意一种单智能体强化学习算法训练多智能体系统。在该类框架中，算法的复杂度随着状态和动作的维度增长呈现维度爆炸 [86]，改问题一般可以通过策略分解（例如 $\boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{o}) = \prod_{i=1}^{n} \pi_i(a^i|o^i)$）或者值分解（例如 $Q_{\text{tot}}(o^1, \cdots, o^n, a^1, \cdots, a^n) = \sum_i Q_i(o^i, a^i)$）来解决。虽然分解类方法可以缓解维度爆炸问题，但是 CTCE 无法评估每个智能体间的相互影响，此时信度分配问题会给学习效率带来较严重的影响。近年来随着多智能体强化学习协作理论的发展，集中式训练分布式执行框架得到广泛关注并取得重大成功。

定义 17 (集中式训练分布式执行) 在 CTDE 框架下，在训练阶段，智能体通过拿到其他智能体的信息甚至是全局信息以优化自己的局部策略：$\pi_i : \Omega^i \to \Delta(\mathcal{A}^i)$. 在分布式执行过程中，智能体仅使用自己的局部信息进行决策。

CTDE 框架近年来在多智能体系统中得到广泛应用，尤其在某些复杂的场景下，更是比前两种范式有更明显的优势。在训练过程中，拿到全局信息，可以缓解非稳态性；在执行过程中，智能体可以直接基于局部信息按策略执行动作。尽管 CTDE 取得了较好的结果，但是目前实验发现，CTDE 在处理异质多智能体（智能体的状态或者动作空间不一致）的时候，往往表现不佳，为了解决这些问题，可以通过技能学习或者通过先分组，再采用局部 CTDE 的方式进行训练，进一步，一些研究工作也从其他方面对集中式与分布式进行讨论 [87,88]

2.2.5 多智能体强化学习的难点与挑战

相较于单智能体强化学习，真实社会中的场景往往更适合建模为多智能体强化学习，然而由于多个智能体的存在，各个智能体的策略同时在更新，往往会带来更多的挑战。本部分对多智能体强化学习中存在的主要挑战包括非稳态 (Non-stationarity)、可扩展性 (Scalability)、部分可观测 (Partial Observability) 及目前存在的解决方法进行介绍。

在单智能体系统中，智能体只需考虑本身与环境的交互即可，此时，环境的转移是固定的。然而在多智能体系统中，环境中多个智能体同时进行策略更新，这样带来的结果是，站在特定智能体的视角，将其他智能体视作环境的一部分，这样的环境就是非稳态 [21] 的，是一个动态目标过程。

定义 18 (非稳态过程) 多智能体中的某一个智能体 i 将会面临一个动态目标过程，也即对任意 $\pi_i \neq \bar{\pi}_i$ 时：

$$P(s'|s, \boldsymbol{a}, \pi_i, \boldsymbol{\pi}_{-i}) \neq P(s'|s, \boldsymbol{a}, \bar{\pi}_i, \boldsymbol{\pi}_{-i}).$$

非稳态表明，随着环境中多个智能体的同时学习与更新，学习不再遵守马尔可夫性。非稳态问题在基于 Q 更新的问题中更严重，尤其是依靠经验回放池更新的一类方法，由于进行 TD 更新需要进行动作采样，在多智能体中，采样未来的联合动作往往难度较大；另一方面，随着所有智能体同时进行更新，经验回放池中的数据可能会过时。解决环境非稳态可以通过对手（队友）建模、对

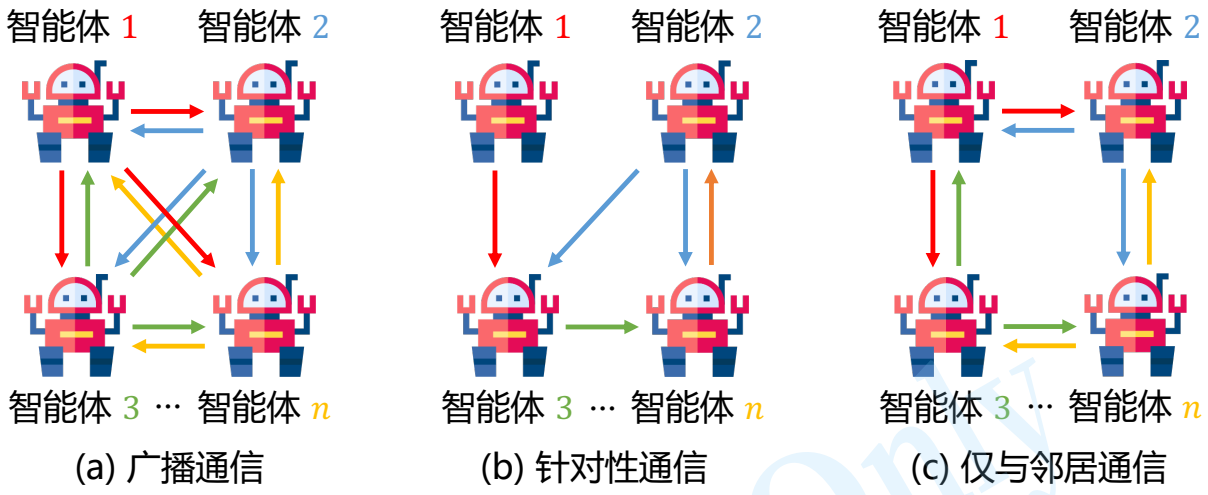

图 7 三种不同的通信拓扑。

回放数据进行重要性采样、进行集中式训练或者采用元强化学习考虑到其他队友的更新等方法。而目前的多智能体方法为解决非稳态问题，一般会考虑环境中其他智能体的动作，带来可扩展性问题。

定义 19（可扩展性） 为了解决非稳态，多智能体往往需要考虑环境中的所有智能体的联合动作，联合动作会随着智能体的数量呈指数上升。

真实场景中的多智能体往往数量较多，比如自动驾驶场景中，存在成百上千甚至更多的智能体，可扩展性是有必要的但也是极具挑战性的。目前的方法包括参数共享（同质智能体共享神经网络，异质智能体间独立训练） [24]；或者通过迁移学习或者课程学习，从少到多进行训练，最终迁移到智能体较多的场景 [60]。

考虑到环境中传感器的限制等因素，智能体往往很难获得全局状态，一般只能拿到部分信息，我们一般把多智能体强化学习建模为 POSG（见定义14），因为智能体只能拿到局部观测信息，从单个智能体的角度，环境不再遵守马尔可夫性，会给训练带来困难。针对上述问题，目前的解决方案包括多智能体通信：智能体之间通过信息传递，缓解多智能体的局部观测问题。通信需要解决的以下问题：与谁通信，通信什么内容，何时通信。对第一问题，目前包括三组拓扑结构如图 7，每个智能体将信息广播（broadcast）给环境中所有的智能体；只发送给特定智能体，把智能体交互建模为网络结构；只与部分邻居进行通信。从通信内容上来看，直接把自身局部信息发送给其他智能体是最直接的方式，然而此种通信方式往往会带来带宽损耗，也存在信息冗余；另外一种方式是发送者对局部信息先进行处理，提取出最有用的信息；针对通信时机，可以在每时每刻都通信，或者在关键时刻进行通信，目前的研究发现，一直通信会带来不必要的信息冗余，如果只在关键时刻通信，通过学到稀疏的通信结构，效率往往会更高。

3 经典环境下的协作多智能体强化学习

3.1 多智能体协作概念

前面我们提到多智能体系统根据任务特性可以划分为协作型，竞争型与混合型这三种设定，其中竞争型一般通过建模为上一章中的零和博弈进行求解，混合型是合作型与竞争型的混合，本部分重点介绍在目前研究较深入的协作型多智能体强化学习算法。

定义 20 (完全协作) 多智能体强化学习中的完全协作指所有智能体共享一个奖赏函数，满足:

$$R^1 = R^2 = \cdots = R^n = R.$$

在协作型多智能体强化学习中，因为智能体共享一个相同的奖赏函数，我们很容易知道智能体间拥有一个相同的回报。由于智能体共享相同的奖赏，如果存在一个中心化的控制器，那么在训练过程中，多智能体强化学习会退化成一个马尔可夫过程，智能体的动作空间会变成一个随机博弈的联合动作空间，其优化目标如下:

$$Q(s_t, \boldsymbol{a}_t) \leftarrow Q(s_t, \boldsymbol{a}_t) + \alpha[r_t + \gamma \max_{\boldsymbol{a}_{t+1} \in \boldsymbol{\mathcal{A}}} Q(s_{t+1}, \boldsymbol{a}_{t+1}) - Q(s_t, \boldsymbol{a}_t)]. \tag{12}$$

在真实世界中，上式难以直接作为多智能体决策的优化目标，这是因为在多智能体系统中智能体往往是独立进行决策的。具体而言，在特定的状态下，可能会存在多组最优联合动作，因为分布式执行的存在，各个智能体可能会偏向于选择不同的最优动作，也就会导致协作行为的涌现。一种最直接的做法是忽略智能体间的协作行为 [89]，我们假设仅存在一个最优均衡点，这样可以直接使用上述公式。早期的分布式 Q 学习算法在全局可观测的假设下，直接忽略智能体间的协作，智能体只更新自己的局部动作价值函数 $Q_i(s, a^i)$。

定义 21 (信度分配问题) 在完全协作性多智能体强化学习中，智能体间共享全局奖赏，如何准确评估某个智能体的动作对整个系统的贡献。

定义 22 (免协作类多智能体协作方法) 假设每个智能体的局部动作价值函数为 $Q_i(s_t, a_t^i)$，我们可以得到其更新为:

$$Q_i(s_t, a_t^i) \leftarrow \max\{Q_i(s_t, a_t^i),\ r_t + \gamma \max_{a_{t+1}^i \in \mathcal{A}^i} Q_i(s_{t+1}, a_{t+1}^i)\}. \tag{13}$$

免协作类多智能体协作方法可以忽略掉奖赏较差的动作，然而一直取最大操作可能会带来过高估的问题，另外一方面，因为动作价值的计算基于联合动作函数，往往扩展性较差。与之相对的是，我们可以考虑协作方式，采用联合学习方式进行策略优化 [90]。

定义 23 (联合动作学习) 假设智能体处于全局可观测的环境中，可以观测到所有队友的状态和动作，每个智能体的策略为 π_i，存在一个全局价值函数 $Q(s, \boldsymbol{a})$，那么对智能体 i，可以评估其局部动作价值为:

$$Q_i(s_t, a_i) =: \sum_{\boldsymbol{a}_{-i} \in \boldsymbol{\mathcal{A}}_{-i}} Q(\langle a_i, \boldsymbol{a}_{-i} \rangle) \prod_{j \neq i} \pi_j(a_j).$$

与此同时，在特殊场景下，可以通过角色分配、协作图、多智能体通信等方式间接实现多智能体协作。与此同时，分布式执行系统往往只能得到局部观测，我们一般将协作型多智能体强化学习建

模为分布式部分可观测马尔可夫决策过程 DEC-POMDP，DEC-POMDP 是 POSG 的一种特殊情况，其中各个智能体拥有相同的奖赏函数。我们发现，如果所有智能体共享奖赏函数和回报，某些智能体不需要对系统作出实质贡献也可以获得奖赏，由此出现了多智能体间的信度分配问题 [22, 91]。

3.2 典型协作方法

3.2.1 基于策略梯度的多智能体协作

在协作型多智能体强化学习中，由于各个智能体间的奖赏与回报相同，使得所有智能体拥有相同的价值函数，我们记全局的动作价值函数为 $Q_{\boldsymbol{\pi}}(s,\boldsymbol{a})$，状态价值函数为 $V_{\boldsymbol{\pi}}(s)$，这里值得注意的是，在多智能体强化学习中某个智能体的动作价值函数与状态价值函数依赖于所有多智能体的策略：

$$\pi_1(a^1|o^1,\theta^1),\pi_2(a^2|o^2,\theta^2),\cdots,\pi_n(a^n|o^n,\theta^n).$$

其中 θ^i 是智能体 i 的策略参数，智能体间的策略差异也主要体现在策略参数的差异上。在策略学习过程中，我们的目的是通过优化策略参数以最大化目标函数：

$$J(\theta^1,\cdots,\theta^n)=E_s[V_{\boldsymbol{\pi}}(s)].$$

所有智能体目标一致，都是通过优化自己的策略参数 θ^i 以最大化目标函数 J，因此上述优化目标也可以写成如下形式：

$$\max_{\theta^1,\cdots,\theta^n} J(\theta^1,\cdots,\theta^n).$$

对某个智能体而言，通过做梯度上升，最大化目标函数：

$$\theta^i\leftarrow\theta^i+\alpha^i\nabla_{\theta^i}J(\theta^1,\cdots,\theta^n).$$

上述的 α^i 是学习率，停止准则是目标函数收敛。上式中我们无法直接计算梯度，一般可以通过价值网络近似策略梯度。

定理 1 (协作型 MARL 的策略梯度定理) 假设存在一个与联合动作无关的基线函数 b，则协作型 MARL 存在如下的策略梯度：

$$\nabla_{\theta^i}J(\theta^1,\cdots,\theta^n)=\mathbb{E}[\nabla_{\theta^i}\log\pi_i(a^i|o^i,\theta^i)\cdot\quad(Q_{\boldsymbol{\pi}}(s,\boldsymbol{a})-b)]. \tag{14}$$

上式中联合动作从联合动作概率中采样：

$$\boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{o},\boldsymbol{\theta})=\pi_1(a^1|o^1;\theta^1)\times\cdots\times\pi_n(a^n|o^n;\theta^n)$$

这里我们重点介绍多智能体深度确定性策略梯度（Multi-Agent Deep Deterministic Policy Gradient，MADDPG）[28]（图 8），MADDPG 的算法伪代码见1，每个智能体对应一个分布式的确定性策略 Actor 模块 $\pi_i(o^i;\theta^i)$，智能体根据 Actor 进行决策；另一方面，每个智能体配备一个集中式的 Critic 网络 $Q_i(s,\boldsymbol{a};\phi^i)$ 用以指导更新优化 Actor 网络。具体而言，MADDPG 会另外维护延迟更新的目标策略 $\pi_i'(o^i;\theta^i_-)$ 用以计算 Critic 的时序差分目标值。从而得到 Critic 的优化目标：

$$\begin{aligned}&y_i^Q=r+\gamma Q_i(s',\boldsymbol{a}'|\phi^i)|_{a'^j=\pi_j'(o'^j;\theta^j_-)},\\&\mathcal{L}(\phi^i)=\frac{1}{2}(y_i^Q-Q_i(s,\boldsymbol{a};\phi^i))^2.\end{aligned} \tag{15}$$

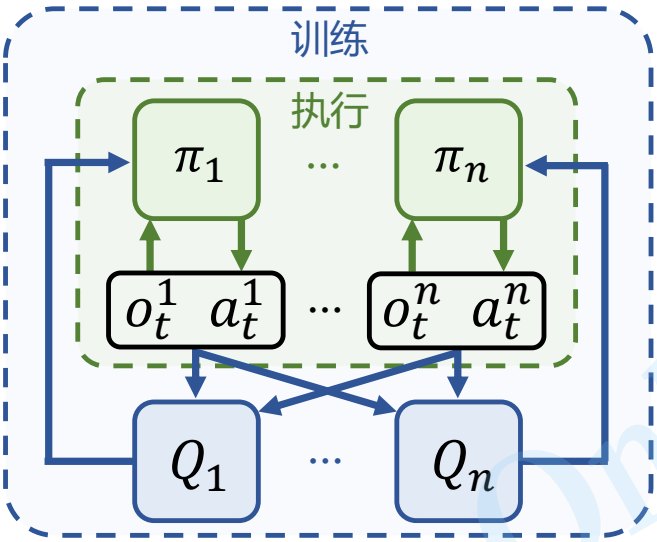

图 8 多智能体深度确定性策略梯度结构图 [28]。

进一步，我们用 Critic 指导 Actor 进行如下更新：

$$\nabla_{\theta^i} J(\theta^i) = \nabla_{\theta^i} \pi_i(o^i;\theta^i) \nabla_{a^i} Q_i(s, \boldsymbol{a};\phi^i)|_{a^j=\pi_j(o^j|\theta^j)} \tag{16}$$

算法 1 MADDPG 算法

for *episode* $= 1$ **to** M **do**
  初始化一个随机的探索噪声 $\mathcal{N}$
  获得初始状态 s
  for $t = 1$ **to** 最大轨迹长度 **do**
    每个智能体依据当前的策略进行探索，得到动作：$a^i = \pi_i(o^i;\theta^i) + \mathcal{N}_t$
    所有智能体执行联合动作 $\boldsymbol{a} = (a^1, \cdots, a^n)$，得到奖赏 r 与新的状态 s'
    将 $(s, \boldsymbol{a}, r, s')$ 存入经验池 $\mathcal{D}$
    $s \leftarrow s'$
    for 智能体 $i = 1$ **to** n **do**
      从经验池 $\mathcal{D}$ 采样样本 $\{s_j, \boldsymbol{a}_j, r_j, s'_j\}_{j=1}^{bs}$
      根据式 15优化 Critic 网络
      根据式 16优化 Actor 网络
    end for
    更新目标神经网络：$\theta^i_- = \tau\theta^i + (1-\tau)\theta^i_-$
  end for
end for

MADDPG 是一个典型的基于 CTDE 框架的算法，Critic 网络的训练需要拿到环境的全局状态信息，在拿不到全局状态信息的环境中，往往需要拿到其他所有智能体的观测信息 o_i，当智能体数

量增多，往往会加大 Critic 网络的训练难度，一般可以采用基于注意力机制 [92] 等方式进行信息聚合，缓解由于数量变化带来的计算代价。另外一方面，MADDPG 可以处理竞争的场景，可能由于隐私等因素，往往无法拿到其他智能体的动作，可以通过对手建模 [59] 实现对其他智能体的动作信息进行估计。我们可以发现，MADDPG 在协作型多智能体强化学习任务中，没有显式考虑信度分配问题，为了显式建模在协作型多智能体系统中各个智能体的贡献（或者应该分得的奖赏），反事实基线多智能体策略梯度算法（Counterfactual Multi-Agent Policy Gradients，COMA）[91] 得到较为广泛的应用，首先定义反事实基线：

定义 24 (协作型多智能体反事实基线) 在协作型多智能体强化学习中，假设存在全局的状态动作价值函数 $Q(s,\boldsymbol{a})$。当其他智能体策略固定的情况下，定义智能体 i 当前动作的优势函数为：

$$A_i(s,\boldsymbol{a}) = Q(s,\boldsymbol{a}) - \sum_{a'^i} \pi^i(a'^i|o^i) Q(s,\langle \boldsymbol{a}^{-i}, a'^i\rangle).$$

上述定义中的 $A_i(s,\boldsymbol{a})$ 通过利用集中式的 Critic 对每个智能体 i 计算出其基线函数，用以计算其策略梯度：

定理 2 (基于 COMA 的多智能体策略梯度) 对于一个基于 Actor-Critic 框架的多智能体强化学习算法，考虑到 $TD(1)$Critic 优化，我们可以得到其策略梯度为：

$$\nabla J = \mathbb{E}\left[\sum_{i=1}^{n} \nabla_{\theta^i} \log \pi_i(a^i|o^i) A_i(s,\boldsymbol{a})\right].$$

3.2.2 值分解的协作方法

虽然基于 CTDE 的多智能体策略梯度方法在部分应用场景展现出较为优异的性能、且针对不同问题，基于注意力机制的方法如 MAAC [92] 能在一定程度上缓解输入维度爆炸的问题、基于反事实基线类方法可以一定程度上解决信度分配问题，但是在复杂的协作场景比如星际争霸微操任务 [33] 中，这类方法的效率较为低下。相反地，基于值函数分解的多智能体强化学习方法有更好的表现 [93]。在接下来的讨论中，我们将值函数的输入从观测 o_t^i 改为轨迹 $\tau_t^i = (o_1, a_1, ..., a_{t-1}, o_t) \in \mathcal{T}^i$，从而一定程度上缓解部分可观测性带来的问题，需要注意的是，在上面所述的基于策略梯度的多智能体协作算法，也可以使用该技巧以提高协作性能。值函数分解的方法大多建立在个体-全局-最大原则（Individual-Global-Max, IGM）的基础上 [82]。

定理 3 (个体-全局-最大原则) 对于一个联合动作价值函数 $Q_{\text{tot}} : \boldsymbol{\mathcal{T}} \times \boldsymbol{\mathcal{A}}$ 以及局部个体动作价值函数 $\{Q_i : \mathcal{T}^i \times \mathcal{A}^i \to R\}_{i=1}^n$，下式成立：

$$\arg\max_{\boldsymbol{a}} Q_{\text{tot}}(\boldsymbol{\tau},\boldsymbol{a}) = \begin{pmatrix} \arg\max_{a^1} Q_1(\tau^1,a^1) \\ \vdots \\ \arg\max_{a^n} Q_n(\tau^n,a^n) \end{pmatrix}.$$

则称 Q_{tot} 与 $\{Q_i\}_{i=1}^n$ 满足个体-全局-最大原则。

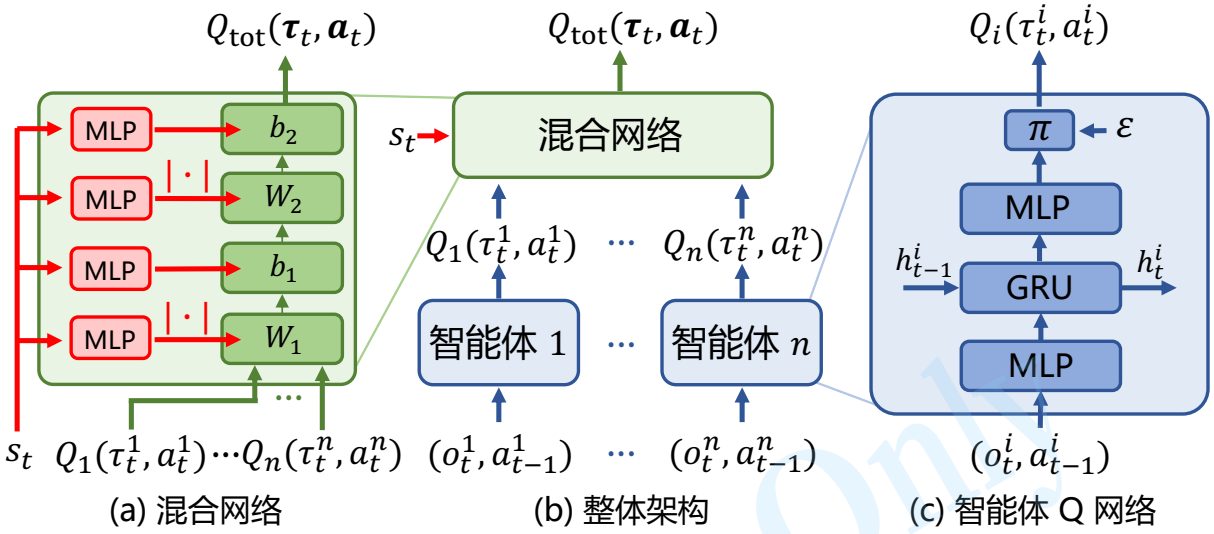

图 9 QMIX 算法框架图。

目前满足 IGM 的典型值分解方法有三种。VDN [30] 用局部 Q 函数的加和表示联合 Q 函数:

$$Q_{\text{tot}}^{\text{VDN}}(\boldsymbol{\tau}, \boldsymbol{a}) = \sum_{i=1}^{n} Q_i\left(\tau^i, a^i\right). \tag{17}$$

这样的值分解方法虽然满足 IGM,但是其网络表达能力不足,在复杂场景下往往表现不佳,QMIX [31] (见图 9) 通过扩展 VDN, 引入混合网络, 结合每个智能体的 Q 值与全局状态计算全局效用, 并进行信度分配。为保证 IGM, 混合网络 (mixing network) 将个体 Q 值与状态作为输入, 得到全局 Q 值并满足一定条件:

$$\begin{aligned} Q_{\text{tot}}^{\text{QMIX}}(\boldsymbol{\tau}, \boldsymbol{a}) =& Mixing(s, Q_1(\tau^1, a^1), ..., Q_n(\tau^n, a^n)), \\ \forall i \in \mathcal{N}, & \frac{\partial Q_{\text{tot}}^{\text{QMIX}}(\boldsymbol{\tau}, \boldsymbol{a})}{\partial Q_i\left(\tau^i, a^i\right)} \geqslant 0. \end{aligned} \tag{18}$$

优化时, 我们计算时序差分损失以进行优化:

$$\mathcal{L}_{\text{QMIX}} = (Q_{\text{tot}}^{\text{QMIX}}(\boldsymbol{\tau}, \boldsymbol{a}) - y_{\text{tot}})^2, \tag{19}$$

其中 $y_{\text{tot}} = r + \gamma \max_{\boldsymbol{a}'} Q_{\text{tot}}^{\text{QMIX},-}(\boldsymbol{\tau}', \boldsymbol{a}')$, 以及 $Q_{\text{tot}}^{\text{QMIX},-}$ 是目标 Q 网络。基于 QMIX 的值分解方法凭借其表达能力与网络结构之间的相对简略, 在各种环境上取得较优的结果, 其伪代码见算法 2。

VDN 和 QMIX 设计所满足的条件是 IGM 性质的充分不必要条件, 导致其在表达能力上仍有一定的欠缺。为了进一步增强神经网络的表达能力, QPLEX [94] 提出一种新的双重对决多智能体 (Duplex Dueling Multi-agent, DDMA) 结构:

$$Q_{\text{tot}}^{\text{QPLEX}}(\boldsymbol{\tau}, \boldsymbol{a}) = V_{\text{tot}}(\boldsymbol{\tau}) + A_{\text{tot}}(\boldsymbol{\tau}, \boldsymbol{a}) = \sum_{i=1}^{n} Q_i\left(\boldsymbol{\tau}, a^i\right) + \sum_{i=1}^{n} \left(\lambda^i(\boldsymbol{\tau}, \boldsymbol{a}) - 1\right) A_i\left(\boldsymbol{\tau}, a^i\right). \tag{20}$$

其中 $\lambda^i(\boldsymbol{\tau}, \boldsymbol{a})$ 是通过多头注意力机制 [95] 计算, 得到的对不同的优势函数进行的信度分配。QPLEX 凭借其表达能力与网络结构, 在多种环境上都取得了较优的结果。

算法 2 QMIX 算法

初始化网络参数 θ，包括混合网络，智能体的局部神经网络，超网络
设置学习率 α，清空经验回放池 $D=\{\}$，$t=0$
while 训练未结束 **do**
  令 step $=0$，s_0 为初始状态
  while s_t 未到终止状态并且 step 小于最大轨迹长度 **do**
    for 每个智能体 i **do**
      $\tau_t^i=\tau_{t-1}^i\cup\{(o_t^i,a_{t-1}^i)\}$
      $\epsilon=epsilon\_schedule(t)$
      $a_t^i=\begin{cases}\arg\max_{a_t^i}Q(\tau_t^i,a_t^i) & w.p.\,1-\epsilon\\ Randint(1,|\mathcal{A}^i|) & w.p.\,\epsilon\end{cases}$
    end for
    获得奖赏函数 r_t 和下一时刻的状态 s_t
    令 $D=D\cup\{(s_t,\boldsymbol{a}_t,r_t,s_{t+1})\}$
    $t=t+1$, $step=step+1$
  end while
  if $|D|\geqslant batch\_size$ **then**
    $b\leftarrow$ 随机从 D 中采样
    for b 中的每条轨迹的每个时间点 t **do**
      通过式 18，计算 $Q_{\text{tot}}^{\text{QMIX}}$ 与 $y_{\text{tot}}^{\text{QMIX}}$
    end for
    通过式 19 计算 $\mathcal{L}_{\text{QMIX}}$ 进行优化。
  end if
  if 更新时间超过设计的阈值 **then**
    更新目标网络
  end if
end while

3.2.3 结合策略梯度与值分解的多智能体协作

除了前述基于策略梯度与值分解类的方法之外，也有部分工作通过集合两者的优势以开发算法，其中可分解的异策略多智能体策略梯度（Off-policy multi-agent decomposed policy gradients, DOP [96] （如图 10 所示），将值分解的思想引入多智能体行动者-评论者框架，并学习一个集中但分解的评论者。该框架将集中式评论者分解为以局部动作为输入的个体评论者的加权线性求和形式。这种分解结构不仅实现了评论者的可扩展学习，还带来了几个好处。它实现了对随机策略的可行异策略评估，减轻了集中式分布式不一致问题，并隐式地学习了高效的多智能体奖赏分配。基于这种分解，DOP 为离散和连续动作空间开发了高效的异策略多智能体分解策略梯度方法。

DOP 的具体贡献包括：(1) 为了解决多智能体的奖赏分配以及学习集中式评论者的可扩展性问

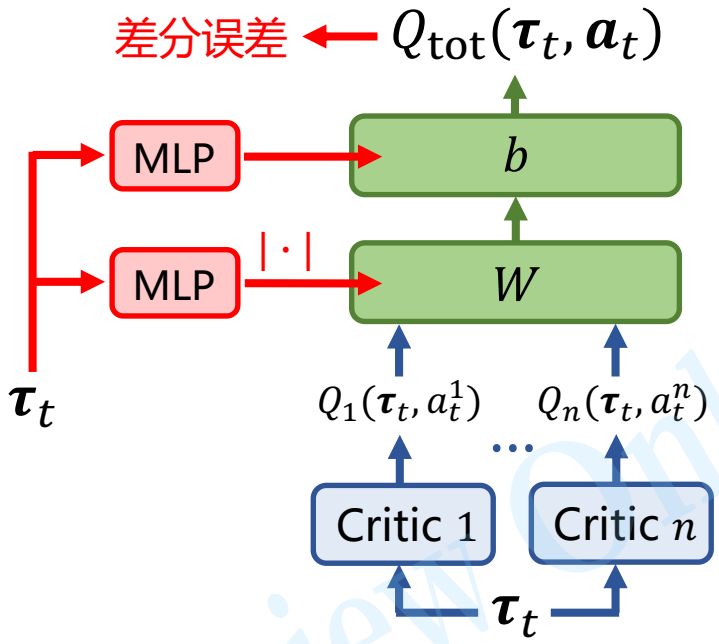

图 10 分解的异策略多智能体策略梯度结构图 [96]。

题，DOP 利用线性值函数分解学习集中化的评论者，这样可以本地推导出每个智能体的基于个体值函数的策略梯度；(2) DOP 提出随机策略梯度和确定性策略梯度多智能体强化学习算法，分别有效执行离散动作任务和连续动作任务；(3) 考虑策略梯度算法的样本效率问题，DOP 结合线性值函数分解和树回溯 (tree backup) 技术，提出可扩展的异策略评论者学习方法，从而提高策略学习的样本效率。除此之外，也有其他的工作从其他方面结合值分解与策略梯度之间的优点，其中 VDAC [97] 直接将策略梯度函数类方法与值分解方法进行结合，提出基于线性求和与基于混合神经网络的方法。IAC [98] 在 MAAC 的基础上通过值分解方法进一步优化智能体间的信度分配，在多个环境上取得令人惊喜的协作效果。FACMAC [99] 提出可分解联合策略梯度的多智能体算法，并对应提出针对连续环境下的多智能体测试环境 MAMuJoCo；FOP [100] 设计一种基于最大熵的值分解与梯度分解方法；另一方面，RIIT [101] 将当前多智能体强化学习中的一些常见技巧整合进当前的协作算法中，并针对性提出设计新的开源算法框架 PyMARL2 [1)]。

3.3 经典环境下的协作多智能体强化学习

除了上一节涉及到的协作能力提升类方法之外，研究者们也从其他方面对协作多智能体进行研究，主要包括高效探索、通信等方面，各方面研究内容及代表工作如表 1 所示。

3.3.1 多智能体协作探索

强化学习方法追求高效地学习到最优策略，其中高质量的训练样本必不可少，探索则在采样过程中扮演着关键角色，是强化学习的一个重要环节 [149]。如在单智能体场景中得到广泛应用的很多探

1) https://github.com/hijkzzz/pymarl2

表 1 经典环境下的协作多智能体强化学习研究内容。

研究方向	核心内容	代表算法	应用与取得成果
算法框架设计	利用多智能体协作理论或设计神经网络提升协作能力	VDN [30], QMIX [31], QPLEX [94], MADDPG [28], MAPPO [29], HAPPO [102], DOP [96], MAT [32]	在多种典型任务场景如 SMAC [33], GRF [32] 等环境上取得不错的协作效果, 展现出巨大潜力
协作探索	设计机制以高效探索环境获得最优协作模式, 与此同时收集高效的经验轨迹以训练策略找到最优解	MAVEN [103], EITI(EDTI) [104], EMC [105], CMAE [106], Uneven [107], SMMAE [108]	在复杂任务场景下显著提升协作效果, 在稀疏奖赏等场景下解决协作能力过低的问题
多智能体通信	设计方法促进智能体间的信息共享, 解决局部可观测等问题, 专注于何时与哪个 (些) 队友交换何种信息	DIAL [109], VBC [110], I2C [111], TarMAC [112], MAIC [113], MASIA [114]	在局部可观测任务场景或需要强协作场景可有效提升协作能力
智能体建模	开发技术赋予智能体推断环境中其他智能体 (实体) 的动作、目标和信念的能力, 促进系统协作能力的提升	ToMnet [115], OMDDPG [116], LIAM [117], LILI [118], MBOM [119], MACC [120]	可以显著改善由于其他智能体的存在带来的环境非稳态问题, 可以在交互关系强与需要强协作的场景下改善协作性能
策略模仿	智能体从给定的轨迹或者示例样本中学习协作策略以完成任务	MAGAIL [121], MA-AIRL [122], CoDAIL [123], DM^2 [124]	实现仅从示例数据进行策略学习的目标
基于模型类方法	从数据中学习世界模型, 多智能体在所学的模型中学习数据以避免与环境直接交互, 提升样本效率	MAMBPO [125], AORPO [126], MBVD [127], MAMBA [128], VDFD [129]	借助成功的模型学习方法或开发针对多智能体系统的方法, 可以显著提升系统的样本利用效率与复杂场景下的协作效能

表 2 经典环境下的协作多智能体强化学习研究内容（续）。

研究方向	核心内容	代表算法	应用与取得成果
动作分层学习	将复杂问题分解成多个子问题, 分别解决子问题进而实现对原始复杂问题求解	FHM [130], HSD [131], RODE [132], ALMA [133], HAVEN [134], ODIS [34]	在多类任务场景下显著提升多智能体系统的协作效率
拓扑结构学习	建模多智能体间的交互关系, 利用如协作图及其他方式刻画智能体间的交互关系	CG [135], DCG [136], DICG [137], MAGIC [138], ATOC [139], CASEC [140]	显（隐）式刻画智能体间的关系, 在复杂场景下可减小系统联合动作空间, 提升协作性能
其他方面	研究可解释、理论分析、社会困境、大规模、延时奖赏等领域	Na2q [141], ACE [142], CM3 [143], MAHHQN [144], 文献 [145~148]	进一步完备协作多智能体强化学习的研究

索技术一样, 多智能体强化学习的探索也在多方面得到关注。原始的多智能体强化学习算法如 QMIX 与 MADDPG 没有专门针对多智能体任务进行特殊探索设计, 在部分复杂场景下表现不佳, 由此针对多智能体探索的工作相继出现。MAVEN [103] 引入一个隐变量并最大化该隐变量与所产生轨迹之间的互信息以改善基于值分解类方法如 QMIX 等的探索能力不足的问题。EITI 与 EDTI [104] 通过考虑智能体间的交互以设计基于协作探索的方法, 在多类环境上表现出较好的协作性能。CMAE [106] 设计受限空间选择来鼓励多智能体探索价值更高的区域以提升协作效率。EMC [105] 将单智能体中基于动力的探索方法扩展到多智能体领域并且提出基于情景记忆的方法以存储高价值轨迹进而促进探索。Uneven [107] 通过同时学习多组任务以得到一个多智能体通用继承性特征来提升智能体在新任务上的探索效率, 达到解决多智能体中的相对过拟合问题的目的。SMMAE [108] 通过均衡多智能体系统中的个体探索与团队探索的关系以提升协作效能, 在复杂任务场景下取得令人满意的效果。ADER [150] 最近被开发出来以研究如何平衡多智能体协作任务中的探索与利用问题。

除此之外, 也有一些其他方法从其他方面对多智能体探索进行研究, 包括多智能体内部奖赏探索 [151]、分布式场景下的多智能体探索 [152, 153]、基于信息获取的探索方法 [154]、未知初始化条件下的协作探索 [155]、多智能体多臂赌博机下的探索问题 [156]、多智能体竞争场景下的探索与利用均衡 [157]、基于组成结构的探索方法 [158] 、乐观探索 [159] 与奖赏稀疏场景下的探索 [160] 等。虽然以上方法从各方面对多智能体探索方法进行研究, 并且在一些测试环境上取得不错的探索效果, 如何开发针对各类探索算法的测试环境与提出有效且完备的多智能体探索理论等都是未来值得研究的课题与方向之一。

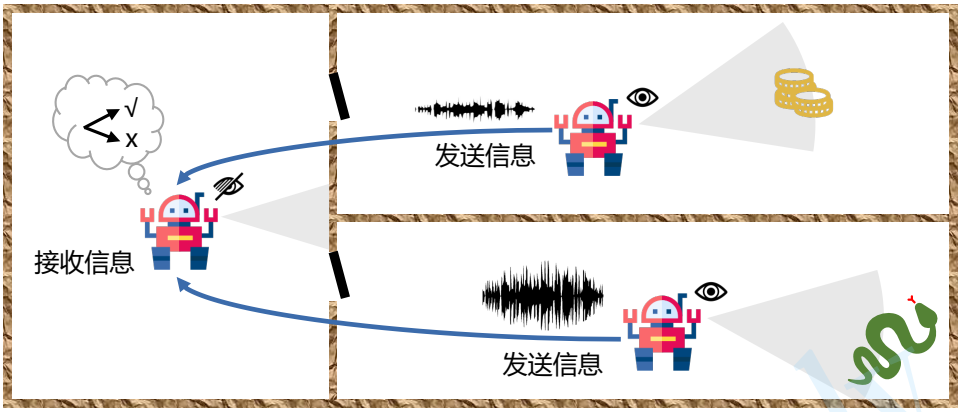

图 11 多智能体通信示意图。

3.3.2 多智能体通信

通信在多智能体系统中扮演着重要的角色，系统内部的智能体基于通信技术分享经验、意图、观测等重要信息（图 11），获得对环境和对其他智能体更清晰的认知，最终更好地协作并完成相应任务 [20]。智能体通信一般涉及三个要素，在何时与哪个（些）智能体交换何种信息，早期方法的主要研究内容是如何更好地将通信与现有的多智能体强化学习相结合以提高多智能体协作效率，其中 DIAL [109] 是一种简单的通信机制，其智能体向所有队友广播信息，以实现端到端的强化学习训练。然而，该方法仅限于离散信息通信场景。CommNet [161] 提出一种高效的集中式通信结构，其中所有智能体的隐藏层输出被收集并平均以增强局部观测。这一方法由于存在信息语义丢失，限制了其在复杂场景中的性能。VBC [110] 添加一个基于方差的正则化器来消除信息中的噪声成分，实现了较低的通信开销和比其他通信方法更好的性能。然而，VBC 需要将智能体的局部信息的隐变量发送给所有队友，不可避免地导致带宽浪费。为了减轻由于信息冗余导致的本地策略负担，IC3Net [162]、Gated-ACML [163] 和 I2C [111] 等研究工作学习一个门控机制来决定在最适当的时机与最合适的队友进行通信。SMS [164] 将多智能体通信建模为一个合作博弈问题，通过夏普利值评估每条信息对于决策的重要性，裁剪没有正增益的信道。MASIA [114] 引入信息聚合的概念，通过自监督信息模块在信息接收端有效提取信息进行决策；另一方面，该文还首次设计开源的离线通信数据集，在在线和离线场景下多种实验表明所设计的通信结构的高效性与合理性。也有部分工作对通信内容的确定展开研究。TarMAC [112] 在信息发送端通过附加签名产生信息，智能体在信息接收端使用注意力机制计算其局部观测与发送自不同的队友的信息签名的相似度，得到一个有针对性的综合信息，并将该综合信息用于增强智能体的局部策略空间。然而，TarMAC 需要将智能体的信息广播给所有队友，导致了严重的带宽浪费。DAACMP [165] 基于 MADDPG [28] 框架，在其中添加一个双重注意力机制，表明注意力可以提高多智能体系统的协作性能。然而，DAACMP 也缺乏对队友的感知，无法实现定制化的通信目的。NDQ [166] 利用最小化通信范式来设计两个不同信息论正则化项来优化值函数，实现局部可观测场景下的近似值分解算法，在一些需要强通信的协作任务中取得较好协作性能。NDQ 正则化的信息会增大本地策略空间，损害学习过程，并且信息的不确定性限制其在复杂环境中的性能。TMC [167] 应用平滑和动作选择正则化器实现简洁而稳健的通信，将收到的信息作为激励

添加到个体值函数中，这是一种高效且有前景的方法。然而，TMC 缺乏队友建模，同时采用了广播通信方式，妨碍其在复杂场景中的协作性能。MAIC [113] 设计一种基于队友建模的分布式激励性通信，在多类任务场景中取得了优异的通信效能，最近也有部分工作如 CroMAC [168]、AME [169]、$\mathcal{R}$-MACRL [49] 与 MA3C [170] 等工作关注于通信策略在部署过程中的健壮性，以获得在策略部署过程中可能面临的通信干扰问题。虽然当前方法在多种任务场景中取得了一定的效果，当前多智能体环境与人类环境存在天然的鸿沟，也就是多智能体所处世界与人类社会世界并不相通，如何通过使用如大语言模型技术等打通两者的鸿沟 [171]，促进人机交互，以及对通信内容进行可解释性研究及理论分析等都是未来值得研究的内容。

3.3.3 基于智能体建模的协作方法

赋予智能体以推断环境中的其他智能体的动作、目标和信念是多智能体强化学习尤其是协作任务中的重点之一 [59]，通过有效且准确的的建模技术，智能体可以高效协作（图 12）。一种最直接的方式是使用心理学中的心智理论 (Theory Of Mind，TOM) 到多智能体系统重以促进智能体协作，其中 ToMnet [115] 及后续的部分工作如 ToM2C [172]、CTH [173]、文献 [174] 等将人类心理学理论引入到多智能体任务中，在一些任务中起到队友建模的作用，可以提升多智能体间的协作能力，更多关于 TOM 的进展可以参见 [175]。

另一方面，其他部分工作通过其他技术去尝试对其他队友进行建模，OMDDPG [116] 运用变分自编码器 (Variational Autoencoders, VAEs)，借助智能体本地信息去推断其他智能体的行为；进一步，LIAM [117] 进一步将该技术扩展到局部可观测条件下，并开发一种高效的队友建模技术。LINDA [176] 通过对队友感知以缓解局部可观测问题，在多种任务场景下协助提升多智能体系统的协作能力。MACC [120] 利用局部信息学习子任务表征并用于增广智能体策略。MAIC [113] 通过局部观测信息预测队友行为并生成激励性消息以促进协作。LILI [118] 考虑机器人场景下的智能体建模问题，开发基于智能体高层策略表征学习的算法以促进协作。SILI [177] 在 LILI 的基础上进一步考虑学习一个平滑的表征以建模其他智能体。文献 [178] 赋予机器人理解自己的行为，并开发技术以更好的帮助人类，达到快速价值对齐的目的。

除了前述所涉及的建模技术之外，SOM [179] 让智能体学习通过使用自己的策略去预测队友的行为，并基于此在线更新自己对于其他队友的未知的目标信念。MBOM [119] 开发基于模型学习的对手建模技术。文献 [180] 在智能体策略学习的同时，将对其他智能体的行为建模作为附加的目标以优化策略学习。DPIQN 与 DRPIQN [181] 考虑建模有策略变化的多智能体任务场景，提出使用策略特征优化器以优化策略学习。ATT-MADDPG [182] 通过基于注意力机制的多智能体版本 DDPG [183] 算法以得到一个联合队友表征来评估队友的策略，以促进多智能体协作。文献 [184] 及其后续工作考虑通过诸如概率推断技术等解决队友建模中的循环推断问题。文献 [185] 提出 TeamReg 与 CoachReg 以相应评估合作团队的动作选择以及更好地进行队友协作。前述工作在多种任务场景下表现出不错的协作效果，然而在这些工作中，智能体需要对环境中的其他智能体或者实体进行建模，当处理大规模智能体系统或实体较多时，使用上述方法会极大增加计算复杂度，如何开发诸如分布式分组等技术手段等以提升其计算效率等，是未来的研究方向之一。

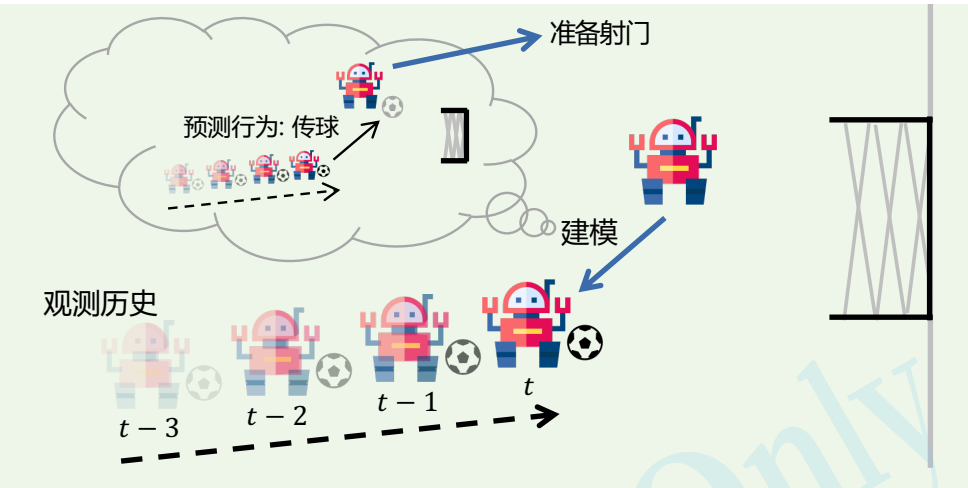

图 12 多智能体建模示意图。

3.3.4 多智能体策略模仿

模仿学习 (Imitation Learning, IL) [186] 指从专家示例中学习，是一种让智能体（机器人）模仿人类专家进行智能决策的方法。由于可以从少量专家轨迹中学到一个具备强适应能力的多智能体策略，多智能体模仿学习最近获得不少进展。其中，文献 [187] 提出从示例数据中学习一个隐空间的概率协作模型，然后用行为克隆技术基于所学的表征与智能体的局部观测输出一组模仿后的策略。文献 [188] 推广前述论文的核心观点，使用分层的结构为每个智能体学习一个高层的意图模型。进一步，文献 [189] 使用连接结构 [190] 以显式地建模智能体间的合作关系。除了行为克隆之外，部分工作也研究多智能体逆强化学习，MAGAIL [121] 将单智能体中成熟的对抗生成技术扩展到多智能体系统中。MA-AIRL [122] 学习开发一种可以应对高维状态空间与未知动态环境的高效和可扩展性强的算法。文献 [191] 根据前述所学的表征空间学习一个奖赏函数以表征多智能体的行为。CoDAIL [123] 开发一种基于相关策略的分布式对抗模仿学习，该方法支持分布式训练与执行。DM^2 [124] 最近开发一种分布匹配的分布式多智能体模仿学习方法。MIFQ [192] 研究一种可分解的逆向多智能体协作算法，在多种场景下展现出优异的协作效果。

另一方面，也有工作从其他方面对多智能体模型学习进行研究，如将多智能体模仿学习用于驾驶模拟 [193]、多智能体逆强化学习中的策略研究 [194]、基于非演示数据的多智能体逆强化学习 [195]、用于平均场博弈问题的对抗强化学习 [196]、用于分布式受限逆强化学习的多智能体系统设计 [197]、非同步多智能体模仿学习 [198] 等。前述工作在从多方面展开研究并且取得一些进展，而考虑到真实多智能体场景的挑战，如何在大规模场景、异质演示数据或次优演示数据场景下学到一个多智能体模仿策略、将多智能体模仿学习应用到真实场景与无人驾驶等，都是未来值得研究的课题与方向。

3.3.5 基于模型的多智能体协作

强化学习由于需要与环境持续交互以进行策略优化，大部分算法是无模型的（model-free），即智能体仅使用与环境交互的样本用于策略优化，往往存在样本利用率低下的问题。基于模型的强化学习是一种被认为具有较高样本利用率的方法，一般的基于模型的方法通常需要学习环境的状态转

移函数（通常也包括奖赏函数），即环境模型，然后基于环境模型的知识去进行策略优化 [199, 200]，可以显著提升样本效率。

近些年来，基于模型的多智能体强化学习也逐步得到关注并取得一定的进展 [35]。早期的一些工作尝试将单智能体中成功应用的技术扩展到多智能体场景，MAMBPO [125] 将单智能体中成功应用的方法 MBPO [201] 扩展到多智能体任务场景，构建基于 CTDE 范式与世界模型的多智能体强化学习，可以一定程度提升系统的采样效率。CTRL [202] 与 AORPO [126] 进一步将对手建模技术引入到模型学习中以获取自适应的轨迹数据以增广原始数据。MARCO [203] 学习一个多智能体世界模型并用其学习一个集中式的探索策略以在高不确定区域收集更多数据增广数据以学习多智能体策略。也有部分技术将在单智能体中广泛应用的 Dreamer [204] 技术扩展到多智能体任务中，MBVD [127] 学习一个基于值分解的隐式的世界模型，在该世界模型中，智能体可以在隐空间中对状态价值进行评估，以使得其拥有 “先见之明”。MAMBA [128] 在合作任务场景中通过基于模型的强化学习来进一步加强集中式训练过程，该文发现在训练过程中可以通过世界模型进行算法优化，在执行过程中可以通过通信搭建一个新的相应的世界模型。VDFD [129] 则通过开发学习解耦世界模型的技术，极大提高多智能体强化学习的样本利用率。

另外一方面，研究者也从其他角度探索多智能体系统中世界模型的学习，比如分布式机器人控制中的防撞 [205]、安全效率提升 [206]、智能体交互建模 [207]、分布式网络系统优化 [208]、基于模型的对手建模 [119]、基于模型学习的多智能体通信效率提升 [209~211]、基于模型的平均场多智能体强化学习 [212]、高效多智能体模型学习 [213]、模型学习以提升离线多智能体强化学习的学习效率 [214] 与基于模型多智能体强化学习的应用 [215] 等。虽然以上方法在一些方面取得一定的成果，但是随着智能体数量变多带来的维度爆炸问题以及分布式执行带来的部分可观测问题，往往阻碍基于模型的多智能体强化学习算法的发展，如何开发可以针对多智能体特色的高效的方法，利用诸如子集选择等手段 [216] 等是后续值得研究的课题与方向之一。

3.3.6 多智能体分层强化学习与技能学习

强化学习在复杂场景可能面临维度灾难问题，为了解决维度灾难，研究者提出了分层强化学习 (hierarchical reinforcement learning, HRL) [217]。HRL 的主要目标是将复杂的问题分解成多个小问题，分别解决小问题从而达到解决原问题的目的。近年来，多智能体分层强化学习也取得一些进展 [218]，FHM [130] 将封建领主结构引入多智能体强化学习过程中，该方法主要针对智能体有不同任务目标的场景，无法适用于共有目标下的协作任务场景。为了解决多智能体任务中的稀疏与奖赏延迟问题，文献 [219] 提出分层版本的 QMIX 算法与分层通信结构，该方法在多个任务场景下取得优异的协作效果，然而该方法需要手动设计高层动作。HSD [131] 通过一个上层宏策略输出相应的技能，该方法的整体结构主要通过监督学习进行训练。RODE [132] 学习动作表征，并通过分层的方式选择动作以提升算法在多种场景下的协作性能。VAST [220] 通过分层解决大规模多智能体协作任务中的效率低下问题。ALMA [133] 充分发掘多智能体任务中的任务结构，并相应提出上层子任务分解策略与下次智能体行动策略。HAVEN [134] 提出一种智能体间和智能体内的双向分层结构以进一步提升多智能体的协作效率。

除前述文章之外，在单智能体强化学习中部分基于技能的工作取得一定进展 [221]，也在多智能

体强化学习中得到应用。文献 [222] 考虑如何将技能整合到多智能体系统中也提升系统性能，HSL [223] 设计方法从多种任务中提取出异质且足够多样的多智能体技能以供下游任务学习，ODIS [34] 则关注于从离线数据集中提取技能，MASD [224] 通过最大化潜在技能分布与所有智能体状态与技能组合之间的互信息，期望发现有益于协作的技能。SPC [225] 关注于多智能体强化学习中的技能种群的自动课程学习过程，文献 [226] 也关注于如何通过学习扩充多智能体技能等。前述所讲方法在一定程度上学习或者使用了多智能体中的技能，然而技能的表现目前仍然缺乏可解释性，如何通过诸如自然语言打通多智能体与人类世界的边界 [227]，使得学到的技能具备可解释性，是后续值得研究的方向与话题之一。

3.3.7 协作多智能体强化学习拓扑结构学习

智能体间的交互是多智能体问题研究中的重点之一，协作图是一种可以通过将多智能体值函数以图的方式进行分解以显式刻画智能体间的交互关系的方法 [135, 228]，最近几年在协作多智能体强化学习中得到广泛应用。在基于协作图的多智能体强化学习任务中，一般可以用节点表示智能体，边（超边）表示有连接关系的智能体间构建在联合动作观测空间上的收益函数，一个协作图可以表示多智能体间价值分解的高阶形式。一般可以通过利用如分布式受限优化 (Distributed Constraint OPtimization, DCOP) 算法 [229] 等寻找具有最大值的动作选择，智能体可以通过连接的边进行多轮信息交换。DCG [136] 引入深度学习的一些技巧将协作图扩充到高维状态动作空间中，在复杂的任务场景如 SMAC 上可以改善多智能体间的相对过泛化问题，DCG 主要专注于在提前训练好的静态和稠密拓扑结构上，在动态环境中可扩展性较差、需要稠密且低效的通信方式。

在协作图问题中的核心问题是如何学到一个动态且稀疏，可以满足智能体动作选择的图结构，该类问题在多智能体学习领域一直得到广泛关注。其中稀疏协作 Q 函数学习 [230] 尝试为价值函数学习稀疏的图结构，该方法学到的图结构是静态并且需要大量的先验知识。文献 [231] 提出为每个智能体学习最小化的动态图集合，该方法存在的问题是计算复杂度随着智能体邻居数量的增多呈指数上升。文献 [232] 通过研究一些图结构的表达能力，但是主要专注于随机拓扑结构与无状态问题。文献 [140] 提出新的协作图测试环境 MACO，并且提出基于上下文感知的稀疏协作图算法 CASEC，该算法通过学习导致表征以有有效减少由于图构建过程中带来的收益函数评估误差，并且在多个环境上取得较好效果。后续工作也从多方面提升协作提，如开发非线性协作图结构 [233]、自组织多项式协作图结构 [234] 等。

另一方面，一些方法也通过利用诸如注意力机制等技术去寻找多智能体间的隐式图结构，通过借助注意力机制去剪断部分不需要的连接结构。如 ATOC [139] 通过注意力机制学习最必要的通信结构，DICG [137] 通过注意力机制学习隐式的协作提结构，MAGIC [138] 以及通过图注意力机制提升多智能体系统的通信效率与组队。以上方法虽然可以在一定程度上得到或者改善多智能体系统交互之间的拓扑结构，但是目前这些方法一般仅能在智能体较少的规模下取得进展，如何在大规模且存在强交互的场景下得到最优拓扑结构，是未来值得研究的方向之一 [235]。

3.3.8 其他方面

除了前文所介绍的目前研究较充分的课题之外，包括多智能体协作可解释性 [141]、开发具备决策先后顺序的多智能体协作算法 [32,142,236,237]、多智能体协作理论分析 [22,238,239]、多智能体多任务多阶段协作学习 [143]、多智能体系统社会困境问题 [240]、异步多智能体协作 [241,242]、大规模多智能体协作 [243]、延时奖赏下的多智能体协作 [244]、混合动作空间下的多智能体协作 [144]、多智能体因果关系发现 [63]、多智能体协作中的课程学习 [145]、协作多智能体强化学习中的教师与学生关系研究 [146]、多智能体中的公平性研究 [147] 与基于实体关系的多智能体协作 [148] 等在内的课题，目前也逐步被探索与挖掘。

3.4 典型测试场景

在算法设计研究的同时，目前有部分工作开发了一系列测试环境对算法进行多方面的验证，典型环境包括星际争霸多智能体挑战（StarCraft Multi-Agent Challenge, SMAC） [33] 及其改进版本 SMACv2 [245]、SMAClite [246]、多智能体粒子世界（Multi-Agent Particle World, MPE） [28]、多智能体版本的 MuJoCo（Multi-Agent MuJoCo, MAMuJoCo） [99]、谷歌足球（Google Research Football, GRF) [247]、大规模多智能体验证环境 MAgent [248] 多智能体离线强化学习测试环境 [249] 与多智能体出租车环境 TaxAI [250] 等。常见多智能体测试环境及其特点见表 3 所示。另外，为了后续工作的方便，也有一些研究者将当前主流的测试环境进行集成并开源，主要包括 Pymarl [33] [2)]、EPyMARL [251][3)]、PettingZoo [252][4)] 、MARLlib [253][5)] 与及第[6)] 等。

3.5 典型应用案例

与此同时，协作多智能体强化学习算法也在多类任务场景中得到广泛应用，包括在游戏领域、工业应用、机器人控制、交叉领域以及国防军事等 [17,275~278]。

早期的多智能体算法主要聚焦于在游戏领域，通过应用强化学习算法以优化策略，AlphaStar [6] 以强化学习为基础，应用诸如种群训练等多智能体学习算法，在星际争霸游戏中控制己方智能体击败地方，获得了惊人的表现 [16,279]。后续很多工作也将多智能体强化学习算法应用在其他游戏任务中，ViVO 团队通过分层强化学习控制王者荣耀中英雄单位进行游戏，在多个场景中取得了良好的对抗效果 [280]，包括 SMAC [33] 在内的诸多测试环境都是基于游戏引擎所开发的。另一方面，除了实时游戏之外，多智能体强化学习在如西洋陆军棋 [281]、中国象棋 [282]、麻将 [283]、斗地主 [284]、足球 [247]、篮球游戏 [261,285] 与捉迷藏 [260] 等方面也都得到人们的关注并且在多种任务场景下取得优异的成果。

另一方面，一些研究者也尝试将多智能体强化学习应用到工业领域，这些研究主要借助多智能体强化学习在问题求解方面的巨大的潜力，将工业问题建模为多智能体强化学习任务。比如文献 [286] 通过将无人驾驶建模为多智能体协作问题，并使用基于协作图的方式促进各无人车之间的协作，发

2) https://github.com/oxwhirl/pymarl

3) https://github.com/uoe-agents/epymarl

4) https://github.com/Farama-Foundation/PettingZoo

5) https://marllib.readthedocs.io/en/latest/index.html

6) http://www.jidiai.cn/environment

表 3 典型多智能体测试环境介绍。

环境名称	是否异质	场景类型	观测空间	动作空间	典型数量	是否通信	问题领域
Matrix Games [90] (1998)	是	混合	离散	离散	2	否	矩阵博弈
MPE [28] (2017)	是	混合	连续	离散	2-6	允许	粒子游戏
MACO [140] (2022)	否	混合	离散	离散	5-15	允许	粒子游戏
GoBigger [254] (2022)	否	混合	连续	连续或离散	4-24	否	粒子游戏
MAgent [248] (2018)	是	混合	连续 + 图像	离散	1000	否	大规模粒子对抗
MARLÖ [255] (2018)	否	混合	连续 + 图像	离散	2-8	否	对抗游戏
DCA [243] (2022)	否	混合	连续	离散	100-300	否	对抗游戏
Pommerman [256] (2018)	否	混合	离散	离散	4	是	炸弹人游戏
SMAC [33] (2019)	是	协作	连续	离散	2-27	否	星际争霸游戏
Hanabi [257] (2019)	否	协作	离散	离散	2-5	是	卡牌游戏
Overcooked [258] (2019)	是	协作	离散	离散	2	否	烹饪游戏
Neural MMO [259] (2019)	否	混合	连续	离散	1-1024	否	多人游戏
Hide-and-Seek [260] (2019)	是	混合	连续	离散	2-6	否	捉迷藏游戏
LBF [251] (2020)	否	协作	离散	离散	2-4	否	食物搜寻游戏
Hallway [166] (2020)	否	协作	离散	离散	2	是	通信走廊游戏
GRF [247] (2019)	否	协作	连续	离散	1-3	否	足球对抗
Fever Basketball [261] (2020)	是	混合	连续	离散	2-6	否	篮球对抗
SUMO [262] (2010)	否	混合	连续	离散	2-6	否	交通控制
Traffic Junction [161] (2016)	否	协作	离散	离散	2-10	是	通信交通调度
CityFlow [263] (2019)	否	协作	连续	离散	1-50+	否	交通控制
MAPF [264] (2019)	是	协作	离散	离散	2-118	否	路径导航
Flatland [265] (2020)	否	协作	连续	离散	>100	否	列车调度
SMARTS [266] (2020)	是	混合	连续 + 图像	连续或离散	3-5	否	无人驾驶
MetaDrive [267] (2021)	否	混合	连续	连续	20-40	否	无人驾驶
MATE [268] (2022)	是	混合	连续	连续或离散	2-100+	是	目标追踪
MARBLER [269] (2023)	是	混合	连续	离散	4-6	允许	交通控制
RWARE [251] (2020)	否	协作	离散	离散	2-4	否	仓库物流
MABIM [270] (2023)	否	混合	连续	连续或离散	500-2000	否	库存管理
MaMo [27] (2022)	是	协作	连续	连续	2-4	否	参数调优
Active Voltage Control [26] (2021)	是	协作	连续	连续	6-38	否	电力控制
MAMuJoCo [99] (2020)	是	协作	连续	连续	2-6	否	机器人控制
Light Aircraft Game [271] (2022)	否	混合	连续	离散	1-2	否	智能空战
MaCa [272] (2020)	是	混合	图像	离散	2	否	智能空战
Gathering [240] (2020)	否	协作	图像	离散	2	否	社会困境
Harvest [273] (2017)	否	混合	图像	离散	3-6	否	社会困境
Safe MAMuJoCo [274] (2023)	是	协作	连续	连续	2-8	否	安全多智能体
Safe MARobosuite [274] (2023)	是	协作	连续	连续	2-8	否	安全多智能体
Safe MAIG [274] (2023)	是	协作	连续	连续	2-12	否	安全多智能体
OG-MARL [249] (2023)	是	混合	连续	连续或离散	2-27	否	离线数据集
MASIA [114] (2023)	是	协作	离散或连续	离散	2-11	否	离线通信数据集

现在多种任务场景下可以取得自动实现车辆调度，其他相关工作也通过多智能体强化学习提升交通信号控制 [287]、无人驾驶 [288]、无人机控制 [289, 290]、无人车控制 [291] 等。除此之外，也有部分研究者从其他方面展开研究，比如通过多智能体强化学习进行电力控制，其中文献 [26] 将主动电力控制建模为 Dec-POMDP 问题，对环境进行开源，并在该环境上对多种多智能体算法进行验证。其他工作也从能源方面展开研究，比如通过多智能体强化学习算法控制风力发电的频率 [292]。也有部分工作在金融领域使用多智能体强化学习 [293~295]，除此之外，一些工作在其他方面展开，MAAB [296] 设计一种基于多智能体强化学习的在线自动竞价框架，以及其他的问题 [297]。除此之外，也有工作通过多智能体对 FPGA 进行时钟同步 [298]，虚拟淘宝 [299] 通过多智能体对抗模仿学习以更好地捕捉用户偏好。也有部分工作通过多智能体强化学习对机器人进行控制 [300~302]，如文献 [303] 通过将协作多智能体强化学习应用到机器人手术的合作协助任务中，表明可以显著提升其任务完成度。

多智能体强化学习也在一些交叉领域得到应用。比如 MA-DAC [27] 将多参数优化问题建模为多智能体强化学习任务，并通过协作多智能体强化学习算法如 QMIX 解决该问题，发现基于多智能体强化学习方法可以显著提升多参数优化能力。MA2ML [304] 通过多智能体强化学习有效处理自动化机器中模块的连接优化学习问题。文献 [305] 研究丰富视觉信息条件下的多智能体强化学习，并设计协作场景下的导航任务。文献 [306] 提出一种基于协作对齐的多摄像头协作系统以解决主动多目标跟踪问题，具体而言，将每一个摄像头当成一个智能体，使用任意一种多智能体算法对其进行控制。文献 [307] 将图像数据增强建模为多智能体问题，该文提出一个更细粒度的自动化数据方法，把一幅图像划分为多个网格，并寻找联合最优增强策略。除此之外，不少工作关注于将多智能体强化学习算法应用到组合优化中，如文献 [308] 研究通过多智能体强化学习优化在线泊车分配，文献 [309] 专注于通过多智能体强化学习框架研究锂离子电池调度问题，文献 [310] 利用多智能体强化学习解决资源抢占环境下的作业调度问题，文献 [311] 探索通过多智能体强化学习研究消小型工厂中的在线调度问题。MARLYC [312] 通过提出一种称为多智能体强化学习偏航控制的新方法以建议控制每台涡轮机的偏航，最终提高农场的总发电量。协作多智能体强化学习在日常生活中也得到一些应用，文献 [313] 通过设计两个协作的机器人对房间内的声学效果进行测试，文献 [314] 通过多智能体系统对日常家居中的能量进行管理，Botzone [315] 研究基于多智能体强化学习的在线智能教育，也有部分工作尝试将多智能体协作技术应用到医疗领域，比如通过多智能体协作对神经元 [316] 或医疗图像进行分割 [317] 等。

除了前文所介绍的应用，多智能体强化学习也在国防应用领域得到一定的探索 [318, 319, 319~323]。文献 [324] 基于 MADDPG 与注意力机制提出一种可以处理队友可变与对手可变场景的多智能体空战任务场景。文献 [325] 提出一种分层多智能体强化方法习 (HMARL) 方法解决异构无人机群协同决策问题，主要针对典型的压制敌方防空 (SEAD) 任务，将其解耦为两个子问题，即高层目标分配 (TA) 子问题和低层协同攻击 (CA) 子问题。文献 [326] 搭建一种基于多智能体强化学习的多无人机协作空战系统，仿真结果表明该文设计的策略学习方法可以获得较高的能量优势并且有效击败多种对手。文献 [327] 提出一种用于空对空作战的分层多智能体强化学习框架以处理多个异构智能体场景。MAHPG [328] 设计了一种基于自博弈对抗训练与分层决策网络的策略梯度多智能体强化学习方法以提升系统空战性能，达到可以学习多种策略的目的。此外，采用层次决策网络来处理复杂的混合动作文献 [329] 设计一种基于两阶段图注意神经网络的机制以捕捉智能空战场景下的智

能体间的关键交互关系，实验表明该方法可以在大规模空战场景显著提升系统的协作能力。

4 开放环境下的协作策略学习

前述内容主要在经典封闭环境，也就是环境中的要素恒定不变的条件下进行讨论的，真实环境中的机器学习算法研究往往需要面对某些要素发生变化的情况，这一特点催生了新的研究领域——开放环境下的机器学习，包括 Open-world Learning [330], Open-environment Learning [38], Open-ended Learning [331], Open-set Learning [332] 等。

4.1 开放环境下的机器学习

传统机器学习一般在经典封闭环境下进行讨论，在这类任务场景下，环境中的重要因素不会发生改变，这里的重要因素在不同的研究领域可以有不同的定义和范围，比如在监督学习中可以是新出现的类别、持续学习的新任务、神经网络输入中的特征增/减、训练（测试）数据分布的变化、任务间学习目标的变化等。根据不同的理解和称谓，开放环境下的机器学习（Open-environment machine learning）[38] 一定程度上与开放学习（Open-ended Learning）存在联系 [331]，与封闭环境下的机器学习相对应，开放环境下的机器学习主要考虑机器学习环境中的重要因素可能发生改变的情况 [38]。过去在该方面有部分工作展开研究，包括类别变化监督学习中的类别增减 [37]、特征变化 [333]、持续学习中的任务变化 [64]、开放环境下的测试环境设置 [334]、博弈问题中的开放性研究 [335] 等。

另外一方面，开放环境下的强化学习近年来也引起不少关注并且在不同方面取得一些进展，不同于监督学习从提供的有标记训练数据中进行学习，或无监督学习对给定的数据内在信息进行分析，强化学习中的智能体被置于一个陌生的环境，需要自主与环境交互，并从交互的结果数据中学习。不同的交互方式会产生不同的数据，因此强化学习中的一个关键难点是从动态的数据分布中学习，而学习的结果还会进一步改变数据的分布。另外，强化学习中的智能体需要考虑考虑所处环境 MDP 的变化，过去的一些方法尝试学习一个具备可信性的强化学习策略 [39]、通过诸如演化学习等方式学习一个具备高泛化能力的策略 [40]、学习技能以面向开放环境 [336]、通过改变奖赏函数以增强泛化能力 [337]、面向开放强化学习的测试环境设计 [338]、基于强化学习的通用开放智能体设计 [339,340]、强化学习中的策略泛化 [42] 等。

4.2 开放环境下的协作多智能体强化学习

前述内容介绍了部分开放环境下的单智能体强化学习相关工作，也有部分研究从特定方面对开放环境下的多智能体强化学习进行描述，其中，开放多智能体系统主要考虑智能体可能在协作过程中加入或离开，导致系统的组成和规模随时间发生变化 [391] 的问题。由于经典的多智能体强化学习算法主要解决包括训练时由于队友策略优化带来的非稳态性，对有效协作模式的探索发掘等问题，虽然能够有效提高样本利用率与协作性能，但是这类方法未将真实世界中多智能体系统与环境因素可能发生变化这一问题纳入考虑范围。以上述开放多智能体系统为例，由于队友行为风格改变，经典的多智能体强化学习算法生成的智能体通过历史信息进行决策，无法及时感知到队友行为风格的改变，自身策略的适应能力有一定滞后性，极大地影响协作性能。在先前的工作中，相关工作主要

表 4 开放环境下的多智能体研究内容。

研究方向	核心内容	代表算法	应用与取得成果
离线学习	将单智能体强化学习中成功应用的技术扩展到多智能体场景或针对性设计多智能体离线方法	ICQ [341], MABCQ [342], SIT [343], ODIS [34], MADT [344], OMAC [345], CFCQL [346]	从收集的静态离线数据中学习策略, 避免与环境交互带来的问题, 实现从大规模及多样性的数据中进行策略的学习目标
策略迁移与泛化	跨任务间学得多智能体策略的迁移与直接泛化, 实现知识重用	LeCTR [347], MAPTF [348], EPC [349], 文献 [350], MATTAR [351]	实现任务任务间的知识重用, 加快在新任务上的学习速度与能力
持续协作	在面对任务或样本以顺序的方式出现的情况下的协作任务学习	文献 [352], MACPro [353], Macop [354]	扩展单智能体已有技术, 在多智能体中处理协作任务流出现情况
演化多智能体强化学习	摸拟生物自然进化过程的启发式随机优化算法, 包括遗传算法、演化策略、粒子群算法等, 赋能多智能体协作	MERL [355], BEHT [356], MCAA [357], EPC [349], ROMANCE [358], MA3C [170]	通过演化算法模拟多智能体策略或生成辅助训练队友或对手帮助多智能体策略训练, 在很多项目场景中得到广泛应用
稳健性研究	考虑系统环境发生变化时的策略学习与执行, 学习可以应对环境噪声、队友变化等情况的稳健策略	R-MADDPG [47], 文献 [46], RAMAO [359], ROMANCE [358], MA3C [168], CroMAC [170]	在环境中状态、观测、动作与通信信道遭受噪声甚至恶意攻击的条件下系统仍具备稳健的协作能力
多目标 (约束) 协作	优化问题中存在多个目标, 需要同时考虑不同目标函数的最优解	MACPO(MAPPO-Lagrangian) [274], CAMA [360], MDBC [361], 文献 [206, 362, 363]	考虑环境中存在的多个约束目标, 在有约束或安全领域等取得进展, 为多智能体协作落地提供基础
风险敏感多智能体协作	使用值分布等手段将环境中的变量数值 (奖赏) 建模为分布, 运用风险函数等评估系统的风险等	DFAC [364], RMIX [365], ROE [159], DRE-MARL [366], DRIMA [367], 文献 [368, 369]	可以在复杂场景下提升协作性能, 在风险敏感场景可以有效感知风险并评估性能

表 5 开放环境下的多智能体研究内容（续）。

研究方向	核心内容	代表算法	应用与取得成果
自组织协作	创建单个自主智能体，使其能够有效、稳健地在特定任务或情境中实现与未知队友的快速协作	文献 [370,371], ODITS [372], OSBG [373], BRDiv [374], L-BRDiv [375], TEAMSTER [376]	赋予单个自主智能体与未知队友快速协作的能力，在多种任务场景下实现临时团队快速协作
零（少）样本协作	设计训练范式以使得多智能体系统在使用少量样本甚至零交互样本的条件下具备与未见队友协作的能力	FCP [377], TrajeDi [378], MAZE [379], CSP [380], LIPO [381], HSP [382], Macop [353], 文献 [52,383]	在多类基准测试环境如 Overcooked 上的结果表明，当前的部分算法可以有效实现与多样未见队友的零样本或少量样本高效协作
人智协同	为人智交互或人机交互提供支持，使得人类参与者与智能体之间更好地协作以完成特定任务	FCP [377], 文献 [384], HSP [382], Latent Offline RL [385], RILI [386], PECAN [387]	在给定的仿真环境或者真实机器人场景等达成一定程度上的人机协作目标
协作大模型	借助通用大模型思想开发决策协作大模型，或者借助当前现有大模型技术促进多智能体协作	MADT [344], MAT [32], MAGENTA [388], MADiff [389], ProAgent [231], SAMA [390]	针对特定任务场景，通过大模型的学习，使得策略具备一定的通用性；此外部分工作借助大语言模型促进系统协作能力

关注于开放情况下的多智能体规划，随之出现很多相关问题设定，如开放式分布式部分可观测马尔可夫决策过程（Open Dec-POMDP）[392]、团队 POMDP（Team-POMDP）[393,394]、轻量级个体 POMDP（I-POMDP-Lite）[371,395] 和 CI-POMDP [396] 等。最近，一些工作开始考虑开放式多智能体强化学习问题。GPL [373] 将开放式即时团队合作问题（Open Ad-hoc Teamwork）形式化为 OSBG （Open Stochastic Bayesian Games），并假设全局可观测性以提高效率，但其在真实世界中难以实现。此外，它使用了基于图神经网络的方法，仅适用于单个可控智能体的设定，导致其扩展到多个可控智能体时比较困难，文献 [397] 最近提出开放多智能系统 OASYS，以对开放多智能体系统进行描述。

虽然前述工作从一些特定方面对开放环境下的智能体强化学习展开研究，但其所关注的内容都相对片面且有一定局限性，缺乏对整个研究领域的全面综述。我们认为一个协作多智能体系统如果希望被应用到复杂的开放真实应用场景，应该具备可以应对环境（状态、动作、奖赏函数等）变化、协作模式（队友、对手）变化、任务以数据流形式出现等问题的能力，主要包括以下方面：

- 策略训练与演进过程中应该至少具备以下能力，包括离线策略学习、策略具备迁移与泛化能力、策略支持持续学习以及系统系统应该具备演化与演进能力；
- 策略在部署过程中具备应对环境因素发生变化的能力，具体而言，在多智能体环境如状态、观测、动作与通信等发生变化下的时具备稳健协作能力；
- 真实环境部署应该考虑的多目标（约束）策略优化、面对真实高动态任务场景时具备风险感知与评估能力；
- 训练好的策略，在部署时，应该具备自组织协作能力，同时应该具有零（少）样本策略适应能力；另一方面，应该支持人智协同，赋予多智能体系统为人类服务的能力；
- 最后，考虑各种多智能体协作任务的差别与相似度，为每一类任务学习一个策略模型往往代价大且浪费资源，策略应该具备诸如 ChatGPT 一样的大模型能力。

基于此，本部分从相应的十一个方面对相关工作进行综述与比较，提出当前研究的主要内容与存在的问题以及未来值得研究的方向。

4.2.1 离线协作多智能体强化学习

离线强化学习 [398, 399] 近来吸引大量研究关注，该类研究专注于一种无需与环境交互的数据驱动训练范式 [399]。先前的部分工作 [400] 主要关注离线学习中的分布偏移问题，并考虑学习行为受限策略以减轻对未见数据估计的外推误差 [401∼403]。离线多智能体强化学习则是一个相对较新且具有前景的研究方向 [404]，它从静态数据集中训练多智能体协作策略。一类离线多智能体强化学习方法尝试从带有策略约束的离线数据中学习策略，其中 ICQ [341] 通过仅信任离线数据有效地缓解多智能体强化学习中的外推误差。MABCQ [342] 引入一个完全分布式的离线多智能体强化学习设定，并利用值偏差和转移归一化等技术来实现高效学习。OMAR [405] 结合一阶策略梯度和零阶优化方法，避免不协调的局部最优解。MADT [344] 利用 Transformer 对序列建模的强大能力，将其与离线和在线多智能体强化学习任务无缝结合起来。文献 [343] 对离线多智能体强化学习进行研究，明确考虑智能体轨迹的多样性，并提出一个名为 Shared Individual Trajectories (SIT) 的新框架。文献 [406] 提出首先训练一个教师策略，该策略有权访问每个智能体的观测、动作和奖赏。在教师策略确定并重组了数据集中的 “好” 行为后，创建单独的学生策略，并将教师策略的特征与智能体之间结构关系通过知识蒸馏赋给学生策略。ODIS [34] 提出一种新颖的离线多智能体强化学习算法，用于从多任务数据中发现协作技能。文献 [249] 最近发布了 Off-the-Grid MARL (OG-MARL) 框架，用于生成离线多智能体强化学习数据集和算法评估，该框架提供了一套初始数据集和基线，并配备一个标准化的评估协议。M3 [407] 创新性地引入多任务和多智能体离线预训练模块思想以学习更高层次可传递的策略表征。OMAC [345] 提出一种基于耦合值分解的离线多智能体强化学习算法，将全局价值函数分解为局部和共享部件，并保持全局状态值和 Q 值函数之间的信用分配一致性。

4.2.2 协作策略迁移与泛化

迁移学习被认为是能够帮助提高强化学习算法样本利用率的重要方法 [408]，旨在不同的任务之间进行知识重用，加快智能体在新任务上的策略学习。多智能体场景下的迁移学习 [60] 研究也得到广泛关注，相比于单智能体而言，除了考虑任务之间的知识重用之外，还有部分研究者针对智能体

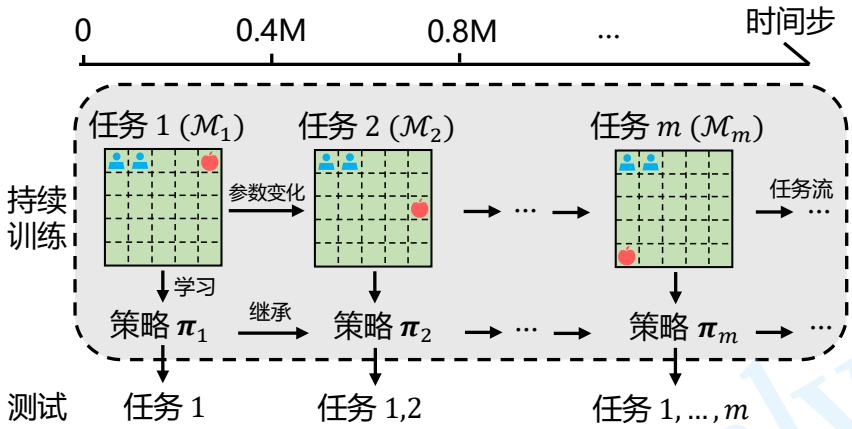

图 13 持续协作示意图。

之间的知识重用进行研究。其基本想法是希望让部分智能体选择性地重用其他智能体的知识，从而帮助多智能体系统整体上实现更好的协作。DVM [409] 将多智能体问题建模为一个多任务学习问题，结合不同任务之间的知识，并且使用一种值匹配机制对这些知识进行蒸馏。LeCTR [347] 在多智能体场景下进行策略教学，让一部分智能体对其他智能体进行引导，从而帮助整体实现更好的策略协作。MAPTF [348] 提出了一种基于选项的策略迁移方法来辅助多智能体协作。

另一方面，多智能体任务之间的策略重用方法强调重用旧任务上的知识经验来辅助新任务的策略学习，关注不同任务之间的知识迁移。相比于单智能体问题而言，不同规模的多智能体任务之间环境输入的维度可能有所不同，这给任务之间的策略迁移带来了阻碍。DyMA-CL [410] 设计一种智能体数量无关的网络结构，并且结合课程学习的思路提出了一系列的迁移机制来加速多智能体协作策略的学习过程。EPC [349] 提出了一种基于演化算法的多智能体课程学习方法，来帮助群体更好地实现复杂场景下协作策略的学习。UPDeT [411] 和 PIT [412] 则借助 transformer 网络的泛化性来处理环境输入维度变化的问题，帮助智能体群体实现高效协作和任务之间的知识迁移。这些多智能体迁移学习的相关工作为任务之间的知识迁移提供了一些启发，但它们没有显式地考虑任务之间的相关性，如何利用任务之间的相关性进行更加高效的知识迁移仍然是非常开放的研究课题。MATTAR [351] 针对协作策略模型难适应新任务的问题，提出了基于任务关系建模的策略迁移方法。文献 [413] 考虑使用边界迁移学习以促进多智能体强化学习。文献 [350] 则进一步关注如何在多智能体强化学习中设计算法以提升其泛化能力。

4.2.3 多智能体强化学习与持续协作

持续学习与增量学习以及终身学习有关，上述学习方式都假定任务或样本以顺序的方式出现 [414]。近年来，持续强化学习 [41, 415] 受到一定关注，在该设定下，智能体面临的挑战是避免灾难性遗忘的同时学会将旧任务的知识迁移到新的任务（又称稳定性-可塑性困境 [416]），同时保持对大量任务的可扩展性。研究者提出多种方法来解决这些挑战，EWC [417] 使用基于 l_2 距离的权重正则化项去约束当前模型参数和过去学习的模型参数之间的差距，该方法需要额外的监督信息来选择特定的 Q 函数头以及为不同的任务场景设置特定的探索方案。CLEAR [418] 是一种与任务无关的

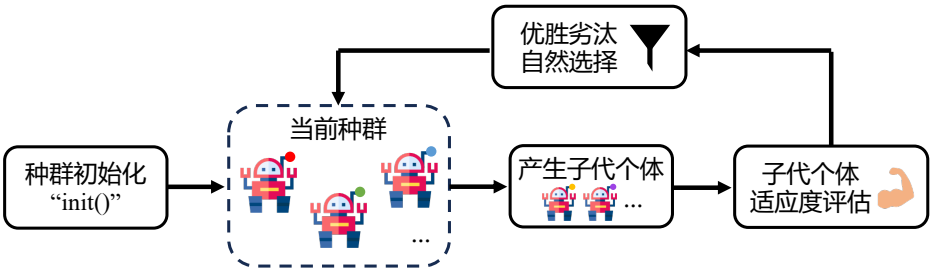

图 14 演化示意图。

持续学习方法，在持续学习过程中不需要任务信息。它存储一个大型的经验回放缓冲区，并通过从过去任务中采样数据来解决遗忘问题。其他方法如 HyperCRL [419] 和 [420]，利用学习的世界模型来增强学习效率。为了解决在任务数量众多的场景中的可扩展性问题，LLIRL [421] 将任务空间分解成子集，并使用中餐厅过程来扩展神经网络，使持续强化学习更加高效。OWL [422] 是一种最近提出的基于多头架构来实现的高效方法。CSP [423] 通过逐步构建策略子空间，在一系列任务上训练强化学习智能体。另外一类可以处理可扩展性问题的手段基于 Packnet [424] 的思想逐次将任务信息编码到神经网络中并对相关任务的网络节点数进行删减。关于多智能体持续学习问题（见图 13），[352] 通过在 Hanabi 基础上引入一个基于多智能体学习的测试平台来研究智能体是否可以与未知智能体协作。然而，它只考虑单模态任务情况。MACPro [353] 提出了一种通过渐进式任务情境化实现多智能体持续协作的方法。它使用共享的特征提取层获得一个任务的特征，但是使用独立的策略输出头，每个策略输出头针对特定类别的任务进行决策。Macop [354] 赋予己方多智能体系统持续协作能力，开发基于非相容队友演化生成与高相容多智能体协作算法对抗训练范式。在队友在回合内与回合间发生切换的多种测试环境中，所提出的方法可以快速捕捉队友身份，较多种对比方法具备更强的适应能力，在面对未知场景下也具有更强的泛化能力。

4.2.4 演化多智能体强化学习

演化算法 [425] 是一类模拟生物自然进化过程的启发式随机优化算法，包括遗传算法、演化策略、粒子群算法等（图 14）。尽管包含诸多变种，其在求解时的主体思路是一致的。首先进行种群初始化，使用随机采样等方式产生若干个体作为初始种群。余下流程可抽象为三个主要步骤的循环，基于当前种群使用交叉、变异等算子产生子代个体；对子代个体的适应度（Fitness）进行评估；根据优胜劣汰的原则淘汰一部分个体，并将余下的个体构成新一代种群。已有研究 [216] 揭示了演化算法在求解子集选择问题上的丰富潜力。演化算法在多智能体领域也得到广泛应用 [426]，比如 α-Rank [427] 及其后续许多工作都使用演化算法对多智能体系统进行了优化和评估。

在协作任务中，演化算法也扮演了重要的角色，文献 [428] 考虑多智能体机器人系统中的分布式配置问题，该文基于模糊系统与进化算法对系统进行优化以提升协作性能。MERL [355] 设计一个分层训练平台，通过两个优化过程分别处理两个目标。一个进化算法通过对团队群体进行神经进化来最大化稀疏的基于团队的目标。与此同时，一个基于梯度的优化器训练策略，只最大化密集的特定智能体的奖赏。BEHT [356] 将多样性高质量目标引入多智能体系统中以解决异质问题，可以有效提升系统的泛化能力。MCAA [357] 及部分后续工作如 [429,430] 考虑通过演化学习改进非对称多

智能体系统。EPC [349] 通过种群演化提升多智能体系统的泛化与迁移能力，ROMANCE [358] 与 MA3C [170] 等方法通过种群演化生成对抗攻击者辅助协作多智能体系统的训练以得到稳健的策略。

4.2.5 稳健协作多智能体强化学习

强化学习中的稳健性研究近年来得到广泛关注并且取得不少进展 [431]，稳健性的研究对象主要包括对智能体不同方面的扰动，例如状态、奖赏、动作等。其中一类方法引入辅助对抗攻击者，通过对待使用策略和对手系统交替的对抗训练来实现稳健性 [432∼435]。其他方法通过在损失函数中设计适当的正则化项来提高稳健性 [436∼438]。相比于对抗训练，这类方法可以有效高了样本利用率。然而，上述方法缺乏针对噪声程度与策略执行之间的稳健性可验证性保证，针对这一问题，相继衍生出一些可验证的稳健性方法 [361, 439∼441]。

目前多智能体强化学习中关于稳健性的研究开始引起人们的关注 [45]，主要困难在于多智能体相较于单智能体需要考虑更多问题，如智能体之间存在的复杂交互带来的非稳态 [21]、信度分配 [22]、可拓展性 [24] 等。早期的部分工作旨在研究协作策略是否具有稳健性。例如，针对一个用 QMIX 算法训练好的协作策略，文献 [442] 通过强化学习训练一个针对观测的攻击者，之后用训练好的攻击者去攻击多智能体中的某一个智能体，结果表明当前通过多智能体强化学习算法训练出来的协作策略对观测扰动是脆弱的。随后文献 [45] 从多智能体系统的奖赏、状态、动作几方面对当前典型多智能体强化学习算法如 QMIX、MAPPO 进行更全面的稳健性测试，结果进一步佐证多智能体系统在面对攻击情况下的脆弱性，表明稳健性多智能体强化学习研究的必要性与迫切性。针对多智能体强化学习稳健性提升方面，近来也有部分工作取得一些进展，其中部分工作关注于学习稳健合作策略来避免过拟合至特定的队友 [46] 与对手 [443]。类似于单智能体稳健强化学习从 MDP 本身包括状态、奖赏、动作方面进行研究，R-MADDPG [47] 关注于多智能体系统的模型不确定性，构建模型不确定下的稳健纳什均衡的概念，在多个环境下取得最优的稳健效果。针对多智能体系统中部分智能体动作被扰动的问题，文献 [48] 引入一个启发式的规则以及相关均衡理论来学习一个稳健的多智能体协作策略。多智能体通信的稳健性也在最近几年得到了一些关注。文献 [444] 设计一个基于高斯过程的过滤器从嘈杂的信息中提取最有价值的信息内容；文献 [445] 在神经网络层面上研究多智能体通信系统的稳健性；文献 [49] 将多智能体通信建模为一个双人零和博弈问题，并应用 PSRO 技术来学习一个稳健的通信策略。ARTS [446]、RADAR [447] 等工作考虑研究多智能体系统的弹性恢复能力，研究协作多智能体强化学习任务在面对环境变化下的恢复能力。

最近一段时间，针对环境动荡下协作稳健难的问题，通过对抗训练提升协作系统在测试环境发生改变时的稳健能力，文献 [358] 提出演化对抗攻击者辅助的稳健协作算法 ROMANCE，文献 [168] 提出通过种群对抗训练的对抗稳健多智能体通信训练框架 MA3C，文献 [170] 提出多视图信息可验证性的稳健通信方法 CroMAC，文献针对多智能体观测噪声扰动进行研究 [359]，文献 [448] 考虑针对多智能体中的状态攻击学习一个稳健策略。

4.2.6 多目标（约束）协作多智能体强化学习

多目标优化 [449] 是指在优化问题中存在多个目标函数，需要同时考虑各个目标的最优解。在多目标优化问题中，不同目标函数之间可能存在相互矛盾的情况，即改善其中一个目标函数可能会导

致另一个目标函数的恶化。因此，需要在不同目标函数之间进行权衡，找到平衡点或者称之为帕累托最优解集。而在多目标强化学习中，类似的，智能体需要根据目标重要性进行权衡，以学习得到对应的帕累托最优策略 [450~453]。在多智能体强化学习中，也有部分工作将多目标学习问题引入进来，一般建模为多目标多智能体系统 (Multiobjective multi-agent systems, MOMAS) [454]，其中不同目标下的奖赏函数可能相互冲突。文献 [455] 考虑多智能体系统中的个体偏好与共享目标之间的关系，将其建模多目标问题，结果表明混合处理方式比单纯考虑单个目标取得更优效果。文献 [456] 探讨了沟通和承诺如何帮助多智能体在具有挑战性的环境中学习到合适的策略。文献 [457] 考虑一般博弈问题中的基于多目标的对手建模问题，通过多目标加速策略学习。

另一方面,最近部分工作关注于单智能体安全强化学习 [458,459] 与多智能体安全强化学习 [274],这类多约束问题在强化学习设定中可建模为受限马尔科夫决策过程 (Constrained MDP, CMDP)。在多智能体强化学习中，文献 [274] 提出针对多智能体任务的测试环境 safe MAMuJoCo、Safe MARobosuite 与 Safe MAIG，并针对提出安全多智能体强化学习算法 MACPO 与 MAPPO-Lagrangian。文献 [460] 研究了在受限制的马尔可夫博弈中进行在线安全的多智能体强化学习,其中智能体通过在对期望总奖赏进行约束的情况下最大化其期望总效用来进行竞争。文献 [461] 研究了安全的多智能体强化学习，其中智能体试图在满足自身安全约束的同时，共同最大化局部目标的总和。CAMA [360] 研究多智能体协作中的安全问题。文献 [462] 考虑状态存在扰动下的多智能体稳健与安全问题。另外一方面，也有部分工作考虑基于屏障保护的安全多智能体强化学习 [206, 362, 363]，或将多智能体安全与控制技术相结合 [361, 463]。

4.2.7 基于风险的多智能体协作

值分布强化学习 (Distributional Reinforcement Learning, Distributional RL) 近年来在多方面取得重要进展 [464]。经典的基于价值的强化学习方法尝试使用期望值对累积回报进行建模，表示为价值函数 $V(s)$ 或动作价值函数 $Q(s,a)$。而在这个建模过程中，完整的分布信息很大程度上被丢失了，值分布强化学习为解决该问题，对累积回报这一随机变量的分布 $Z(s,a)$ 进行建模。该类方法在多智能体协作任务中也得到应用，为了缓解由于局部可观测带来的环境的随机性，DFAC [364] 将单个智能体的收益函数由确定性变量扩展为随机变量，然后将 QMIX 类算法的混合函数建模为配分混合函数，该算法在多种高难度的协作任务上取得了较为优异的协作效果。进一步，为了缓解多智能体协作任务中的奖赏函数的随机性带来的不确定性，RMIX [365] 借助基于风险的值分布技巧 (Conditional Value at Risk, CVaR) 提升算法的协作能力，该算法创新性地提出基于智能体轨迹相似度评估的风险度评估，理论证明 RMIX 算法的合理性，实验结果也进一步验证算法的有效性。ROE [159] 从另外的角度提出一种基于风险评估的协作动作乐观探索方法，该方法可以针对性地对分布进行采样，可以有效提升多智能体的探索效率。

除了上述所涉及到的基于值分布的多智能体协作算法之外，也有一些工作从其他方面展开，比如基于值分布的奖赏评估 [366]、一般博弈问题中的自适应高效多智能体策略学习 [368]、多智能体学习过程中的风险解耦学习 [367]、基于博弈论的风险管理 [369] 等。虽然所述工作在多种环境上取得了一定成果，但是考虑到真实环境的风险未知，如何在将多智能体策略部署到真实环境中，通过自动识别环境风险并相应调整协作策略是未来值得研究的方向之一。

4.2.8 自组织多智能体协作

自组织队友协作 (Ad hoc teamwork, AHT) [50] 旨在创建一个自主智能体，赋予其能够有效、稳健地与之前未知的队友在进行高效协作的能力 [370]。早期的工作假设需要配合的队友行为对于学习的智能体而言是已知的 [370, 465]。后续相关工作逐步放松该假设，在交互过程中，学习的自主智能体无法提前获知队友的行为。一些工作设计算法通过观测队友的行为去预测相应队友的策略，进而促进其协作过程 [466~469]。另外一些工作试图通过有效的通信手段以进一步提升提升 AHT 中队友之间的协作能力 [470]。以上工作虽然在一定程度上可以提升协作性能，但是这些方法假设合作的队友处于封闭的环境中，即队友在单轨迹内的数量和类型保持不变，基于此，开放 AHT 相应的被提出并研究 [371]，其中 GPL [373] 通过构建图神经网络应对不同时刻队友类型与数量的变化。

早期关于 AHT 的工作一般考虑 AHT 的智能体处于全局可观测的环境中，最近的部分工作考虑将该设定扩展到局部可观测场景，其中 ODITS [372] 设计基于互信息优化的正则项去评估其他队友的行为，训练得到的自主智能体仅用其局部观测即可推测队友的行为。与先前的研究不同， [471] 提出一种在部分可观测情况下解决开放式临时团队合作问题的方法。TEAMSTER [376] 提出一种解耦环境模型和队友行为模型学习的方法。除此之外，部分工作从其他方面展开工作，包括存在攻击者的 AHT 问题 [472]、少样本交互协作 [52]、AHT 的队友生成覆盖问题 [374, 375] 等。

4.2.9 零（少）样本协作

零样本协作 (Zero-shot coordination, ZSC) 是近年来在协作多智能体任务中被提出来的概念，其目标是训练智能体使其能够与未见队友协作 [51]。自博弈（或“自我训练”） [473, 474] 是有效提升协作能力的手段之一，智能体通过与自身进行协作训练不断提升，但是由此生成的智能体可能缺乏与未见队友协作的能力。文献 [475] 进一步规范该问题，添加顺序无关的训练方法以缓解次优性。由于仅与单个队友训练配合可能存在过拟合到特定队友行为风格的问题，其他方法如虚拟博弈 (Fictitious Co-Play, FCP) [377, 476]，或智能体和队友种群的协同演化 [379] 等技术都在该方面取得一定的成果。也有部分工作利用少样本技术来应对多模态场景 [353, 380] 并且取得一定的成效。文献 [477] 最近通过合作队友的动作偏好对当前多种 ZSC 算法的容量进行评估与计算。

除了前述工作之外，ZSC 系列工作还包括多样性度量 [378, 478]、训练范式设计 [377, 379, 475]、等量可交互网络设计 [479]、策略相似性评估下的协作提升 [480]、通用场景下的 ZSC 问题 [481]、基于集成技术提升的 ZSC [481]、人类价值偏好研究 [382]、多样队友生成 [381] 以及异构环境下的策略协同演化 [379] 等。此外，少样本适应在单智能体元强化学习中中被广泛应用 [482~484]，在该设定下，使用一个探索策略在环境中采样 K 条轨迹以获得对下游任务推断的隐变量 z 从而决策。具体而言，在少样本队友协作 (Few-shot Teamwork, FST) [52] 中，在不同任务中训练生成的智能体需要组成新的团队，在未知但相关的任务中，能够适应并协作完成任务。其中 CSP [380] 考虑多模态协作范式，开发一种协作策略与探索策略解耦的少样本协作范式，在策略执行时采集少量样本以寻找最优策略头。文献 [383] 发现当前性能较优的 ZSC 算法在面对不同的学习方法时需要较多的样本去适应新的队友，相应提出一种基于少样本的协作方法，并在 Hanabi 上验证算法的有效性。Macop [354] 考虑回合间协作对象发生变化下的策略适应能力，提出面向任意队友的高相容性协作算法，显著提升协作算法的泛化能力。

4.2.10 人智协同

赋予智能体（机器人）与人类高效协作的能力是人工智能一直以来的目标之一 [485, 486]。其中人智协同 [258] 可以为人智交互（Human-AI Interaction, HAI) [487] 或人机交互（Human-Robot Interaction, HRI) [488] 提供支持，人智协同的目的是为了使得人类参与者与智能体之间更好地协作以完成特定任务，针对不同的人类参与者，可以通过协作多智能体强化学习提升人智协同能力。

与前述提到的 ZSC 问题不同的是，人智协同考虑的合作对象是人类参与者。虽然曾经有研究在部分环境下发现可能不需要人类数据训练智能体模型就能与真实人类合作 [377]，但在人类行为的某些微妙特征对任务产生至关重要影响的场景下，是无法脱离人类数据产生良好协作策略的。对于待训练的智能体而言，可以通过先验偏置去直接编码具有人类行为风格的队友 [475, 489, 490]，也可以不同程度的利用与真实人类交互的数据进行训练。一些方法则同时结合基于先验偏置手工编码人类行为与人类交互数据对智能体进行优化 [491~493]

然而，上述方法都对测试时人类行为的模式做出较强假设，在实际上难以成立。针对这一问题，涌现出一系列学习人类行为模型和计算对其的最佳响应的方法来实现人工智能与人类合作的方法 [384]。在人智协同中，目前有部分方法从其他角度进行研究，包括研究人类有偏好的任务场景 [382]、通过离线数据促进人机协作 [385]、开发技术以实现人智相互配合 [386]、人机协作中的领导与跟随技术开发 [494]、零样本人智协作 [495]、基于贝叶斯优化的人智协作等 [496] 以及人机协作环境设置 [497] 等。虽然以上工作在人智协同方面取得一些成果，然而该方向目前依然存在不少问题与挑战，比如，缺乏便捷且有效的测试环境，大部分工作主要在 Overcooked [258] 上开展，该测试环境存在智能体数量较少且部分场景过于简单的问题；另外一些工作主要在第三方不开源的自制环境比如机械臂上进行验证，如何在未来开发具备多种任务需求且适用人类参与者上手，便于测试的测试环境，以及设计更高效的算法等都是未来值得研究的方向。另一方面，通过人类价值对齐 [498,499] 人类人在环训练 [500] 等，是解决以上问题的可能方案之一。

4.2.11 协作多智能体强化学习与大模型

大模型，尤其是大语言模型 [501]，近年来在多个领域引发广泛关注并且得到应用，目前也有部分工作关注于通用决策大模型的探索 [502, 503] 并得到一定应用。在单智能体强化学习任务中，GATO [504]、DreamerV3 [505]、DT [506] 等研究工作在许多任务场景下取得令人惊喜的效果，这些工作都依托于先有技术如 Transformer [95] 的强大的表达能力 [507]。另一方面，近来的一些工作也尝试学习多智能体通用决策大模型，其中 MADT [344] 通过提供大规模数据集来促进研究，并将其用于在多智能体强化学习环境下探讨 DT 的应用。MAT [32] 研究一种可以有效地将 MARL 转化为单智能体问题的大模型方法，其目标是将智能体的观察序列映射到智能体的最优行动序列，在多个任务场景下表现出比传统方法更优越的性能效果。为了充分利用多智能体环境中的实体信息，文献 [388] 从实体的角度提出了多智能体多游戏实体 MAGENTA，这是对先前的时间序列建模的一个正交研究。MADiff [389] 与 DOM2 [508] 将生成式扩散模型引入到多智能体强化学习以促进系统的协作，在多种场景下取得最优效果。SCT [509] 通过 Transformer 模型加速多智能体在线适应。文献 [510] 构建类人的多智能体环境，“西部世界”，以模拟并测试大规模场景下的多智能体系统。

另一方面，随着以 ChatGPT 为代表的大语言模型的发展，目前也有部分工作尝试通过语言模

型促进多智能体协作，其中 EnDi [511] 尝试使用自然语言以提高多智能体系统的泛化能力。InstructRL [512] 使人类能够通过自然语言指令来获得期望的智能体策略。SAMA [390] 提出基于语义对齐的多智能体协同，该方法通过预训练好的语言提示词自动给多智能体系统分配目标，在多种场景下取得了令人欣喜的协作效果。ProAgent [513] 提出一种可以充分利用大语言模型以进行队友行为预测的高效人机协作框架，在人机协作任务场景下取得最好的协作性能。另一方面，也有部分工作将多智能体协作方法应用在大语言模型能力的提升中 [514~516]。然而由于复杂交互等原因，多智能体通用决策大模型目前还是一个较少涉及的领域，如何在目前已有的多种多智能体任务场景下学到一个通用的多智能体决策大模型，可以在多种场景下进行零样本或者少样本泛化，或者通过微调可以获得一个在未见任务场景的快速学习的策略等，都是极具挑战但是值得研究的方向。

5 总结与展望

本文着眼于协作多智能体强化学习的发展与研究，从当前主流的经典环境到与现实应用贴近的开放环境下的多智能体协作逐层深入。对强化学习、多智能体系统、多智能体强化学习、协作多智能体强化学习等方面进行概括性的介绍，针对不同研究方向进行总结，归纳凝练出经典环境下的多智能体强化学习研究关注点。虽然目前许多基于封闭假设的多智能体强化学习算法在经典环境中取得了优异的协作效果，但是在真实世界的应用仍然有限，这很大程度应该归结于现有方法未对开放环境的特点进行针对性的研究，离真正赋能日常生活还存在较大差距。为克服开放环境下交互繁杂，系统多变，约束众多等障碍，未来的协作多智能体强化学习可以从以下角度开展相关研究，并基于此引发更多对开放环境下多智能体强化学习的关注，让协作多智能体系统更好的被应用在现实环境中，以赋能人类生活。

- **经典封闭环境下的多智能体协作问题解决**。在经典环境下的协作多智能体强化学习算法是走向开放环境的基石，通过提升系统在封闭环境下的协作性能，可以让其具备更广的应用潜力。然而目前封闭环境下的协作多智能体强化学习仍然存在诸多问题有待解决。包括在智能体数量极多的大规模场景下进行高效的策略优化 [517]，如何平衡训练与执行过程中分布式与集中式的关系等，都是未来研究工作需要考虑与解决的。
- **开放环境下的理论分析与体系构建**。相较于封闭环境，开放环境中的条件更严苛，挑战更艰巨。目前部分工作采用启发式规则对机器学习环境开放程度进行设计 [38]。但是针对协作多智能体系统所处环境，如何搭建包括多智能体协作开放性的完备定义，环境开放性程度定义，以及算法的性能边界等在内的一套体系，是未来需要进一步考虑的问题。
- **开放环境下的协作多智能体测试环境搭建**。虽然当前有部分工作从稳健性等方面对开放环境下的协作多智能体强化学习展开研究，但基准测试仍然是在经典封闭环境上进行修改所进行的，仅考虑文章的特有设定，缺乏兼容不同开放性挑战的测试环境与统一的评估标准。如何搭建包含对前述十一个方面进行评估的测试环境，是未来促进开放环境下的协作多智能体研究的极大助力。
- **面向开放环境下的多智能体通用决策大模型**。大模型，尤其是大语言模型 [501] 近年来在多个领域引发广泛关注并且得到应用。部分工作探索通用决策大模型 [503, 518] 并且在一些领域得到应用。然而由于复杂繁杂，实体多变等原因，多智能体通用决策大模型这一研究领域仍然存在较大空白，在未来，研究者需要考虑如何在多种任务场景下学习通用的多智能体决策大模型，以实现零样

本或少样本泛化，或通过微调实现对未知任务领域的快速适应。

• **协作多智能体强化学习应用与落地**。经典环境中多智能体强化学习的高效协作性能显现出极大的应用潜力，然而大部分工作仅限于在模拟器或特定任务场景中进行测试 [519]，与真实的社会应用场景与需求目前还有一段距离。而针对开放环境下的协作多智能体强化学习的研究的主要目的仍然是促进算法应用到人类生活并促进社会进步。在未来，如何将多智能体强化学习算法安全高效的应用在诸如大规模无人驾驶、智慧城市、大规模计算资源调度等领域，都是十分值得探讨的话题。

致谢 感谢秦熔均、张福翔、王铖鹤、李逸尘、薛科、贾乘兴与陈烽等同学提供的宝贵建议与帮助。

参考文献

1 Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018. 1, 4, 7
2 Ben Goertzel and Cassio Pennachin. *Artificial General Intelligence*. Springer, 2007. 1
3 Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 1
4 Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *preprint arXiv:1312.5602*, 2013. 2
5 David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. 2
6 Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019. 2, 30
7 Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Dkebiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *preprint arXiv:1912.06680*, 2019. 2
8 Junjie Li, Sotetsu Koyamada, Qiwei Ye, Guoqing Liu, Chao Wang, Ruihan Yang, Li Zhao, Tao Qin, Tie-Yan Liu, and Hsiao-Wuen Hon. Suphx: Mastering mahjong with deep reinforcement learning. *preprint arXiv:2003.13590*, 2020. 2
9 Yuxi Li. Deep reinforcement learning: An overview. *preprint arXiv:1701.07274*, 2017. 2, 8
10 Ashish Kumar Shakya, Gopinatha Pillai, and Sohom Chakrabarty. Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications*, page 120495, 2023. 2
11 Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *preprint arXiv:2304.01852*, 2023. 2
12 Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023. 2
13 Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan D. Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, Craig Donner, Leslie Fritz, Cristian Galperti, Andrea Huber, James Keeling, Maria Tsimpoukelli, Jackie Kay, Antoine Merle, Jean-Marc Moret, Seb Noury, Federico Pesamosca, David Pfau, Olivier Sauter, Cristian Sommariva, Stefano Coda, Basil Duval, Ambrogio Fasoli, Pushmeet Kohli, Koray Kavukcuoglu, Demis Hassabis, and Martin A. Riedmiller. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nat.*, 602(7897):414–419, 2022. 2
14 Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022. 2
15 Ali Dorri, Salil S Kanhere, and Raja Jurdak. Multi-agent systems: A survey. *IEEE Access*, 6:28573–28593, 2018. 2, 3, 8
16 Yaodong Yang and Jun Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *preprint arXiv:2011.00583*, 2020. 2, 3, 30
17 Afshin Oroojlooy and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023. 2, 3, 30

18 Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2023. 2

19 Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 55(2):895–943, 2022. 2, 10

20 Changxi Zhu, Mehdi Dastani, and Shihan Wang. A survey of multi-agent reinforcement learning with communication. *preprint arXiv:2203.08975*, 2022. 2, 3, 12, 25

21 Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V Albrecht. Dealing with non-stationarity in multi-agent deep reinforcement learning. *preprint arXiv:1906.04737*, 2019. 2, 3, 14, 39

22 Jianhao Wang, Zhizhou Ren, Beining Han, Jianing Ye, and Chongjie Zhang. Towards understanding cooperative multi-agent q-learning with value factorization. In *Advances in Neural Information Processing Systems*, pages 29142–29155, 2021. 2, 17, 30, 39

23 Chongjie Zhang. *Scaling multi-agent learning in complex environments*. PhD thesis, 2011. 2

24 Filippos Christianos, Georgios Papoudakis, Muhammad A Rahman, and Stefano V Albrecht. Scaling multi-agent reinforcement learning with selective parameter sharing. In *Proceedings of the International Conference on Machine Learning*, pages 1989–1998, 2021. 2, 15, 39

25 Guillaume Sartoretti, Justin Kerr, Yunfei Shi, Glenn Wagner, TK Satish Kumar, Sven Koenig, and Howie Choset. Primal: Pathfinding via reinforcement and imitation multi-agent learning. *IEEE Robotics and Automation Letters*, 4(3):2378–2385, 2019. 2

26 Jianhong Wang, Wangkun Xu, Yunjie Gu, Wenbin Song, and Tim C Green. Multi-agent reinforcement learning for active voltage control on power distribution networks. In *Advances in Neural Information Processing Systems*, pages 3271–3284, 2021. 2, 31, 32

27 Ke Xue, Jiacheng Xu, Lei Yuan, Miqing Li, Chao Qian, Zongzhang Zhang, and Yang Yu. Multi-agent dynamic algorithm configuration. In *Advances in Neural Information Processing Systems*, pages 20147–20161, 2022. 2, 31, 32

28 Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pages 6379–6390, 2017. 2, 17, 18, 23, 25, 30, 31

29 Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, pages 24611–24624, 2022. 2, 23

30 Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinícius Flores Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 2085–2087, 2018. 2, 20, 23

31 Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 4295–4304, 2018. 2, 20, 23

32 Muning Wen, Jakub Grudzien Kuba, Runji Lin, Weinan Zhang, Ying Wen, Jun Wang, and Yaodong Yang. Multi-agent reinforcement learning is a sequence modeling problem. In *Advances in Neural Information Processing Systems*, pages 16509–16521, 2022. 2, 23, 30, 35, 42

33 Mikayel Samvelyan, Tabish Rashid, Christian Schröder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob N. Foerster, and Shimon Whiteson. The Starcraft multi-agent challenge. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 2186–2188, 2019. 2, 19, 23, 30, 31

34 Fuxiang Zhang, Chengxing Jia, Yi-Chen Li, Lei Yuan, Yang Yu, and Zongzhang Zhang. Discovering generalizable multi-agent coordination skills from multi-task offline data. In *International Conference on Learning Representations*, 2023. 2, 24, 29, 34, 36

35 Xihuai Wang, Zhicheng Zhang, and Weinan Zhang. Model-based multi-agent reinforcement learning: Recent progress and prospects. *preprint arXiv:2203.10603*, 2022. 2, 3, 28

36 Xueguang Lyu, Yuchen Xiao, Brett Daley, and Christopher Amato. Contrasting centralized and decentralized critics in multi-agent reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 844–852, 2021. 2

37 Jitendra Parmar, Satyendra Chouhan, Vaskar Raychoudhury, and Santosh Rathore. Open-world machine learning: appli-

cations, challenges, and opportunities. *ACM Computing Surveys*, 55(10):1–37, 2023. 3, 33

38 Zhi-Hua Zhou. Open-environment machine learning. *National Science Review*, 9(8), 2022. 3, 33, 43

39 Mengdi Xu, Zuxin Liu, Peide Huang, Wenhao Ding, Zhepeng Cen, Bo Li, and Ding Zhao. Trustworthy reinforcement learning against intrinsic vulnerabilities: Robustness, safety, and generalizability. *preprint arXiv:2209.08025*, 2022. 3, 33

40 Rui Wang, Joel Lehman, Aditya Rawal, Jiale Zhi, Yulun Li, Jeffrey Clune, and Kenneth Stanley. Enhanced poet: Open-ended reinforcement learning through unbounded invention of learning challenges and their solutions. In *Proceedings of the International Conference on Machine Learning*, pages 9940–9951, 2020. 3, 33

41 Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022. 3, 37

42 Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76:201–264, 2023. 3, 33

43 Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. A survey of meta-reinforcement learning. *preprint arXiv:2301.08028*, 2023. 3

44 Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence*, pages 737–744, 2020. 3

45 Jun Guo, Yonghong Chen, Yihang Hao, Zixin Yin, Yin Yu, and Simin Li. Towards comprehensive testing on the robustness of cooperative multi-agent reinforcement learning. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 114–121, 2022. 3, 39

46 Tessa van der Heiden, Christoph Salge, Efstratios Gavves, and Herke van Hoof. Robust multi-agent reinforcement learning with social empowerment for coordination and communication. *preprint arXiv:2012.08255*, 2020. 3, 34, 39

47 Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. In *Advances in Neural Information Processing Systems*, pages 10571–10583, 2020. 3, 34, 39

48 Yizheng Hu, Kun Shao, Dong Li, Jianye Hao, Wulong Liu, Yaodong Yang, Jun Wang, and Zhanxing Zhu. Robust multi-agent reinforcement learning driven by correlated equilibrium, 2021. 3, 39

49 Wanqi Xue, Wei Qiu, Bo An, Zinovi Rabinovich, Svetlana Obraztsova, and Chai Kiat Yeo. Mis-spoke or mis-lead: Achieving robustness in multi-agent communicative reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 1418–1426, 2022. 3, 26, 39

50 Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V Albrecht. A survey of ad hoc teamwork research. In *European Conference on Multi-Agent Systems*, pages 275–293, 2022. 3, 41

51 Johannes Treutlein, Michael Dennis, Caspar Oesterheld, and Jakob Foerster. A new formalism, method and open issues for zero-shot coordination. In *Proceedings of the International Conference on Machine Learning*, pages 10413–10423, 2021. 3, 41

52 Elliot Fosong, Arrasy Rahman, Ignacio Carlucho, and Stefano V Albrecht. Few-shot teamwork. *preprint arXiv:2207.09300*, 2022. 3, 35, 41

53 Yoav Shoham, Rob Powers, and Trond Grenager. Multi-agent reinforcement learning: a critical survey. Technical report, Citeseer, 2003. 3

54 Yoav Shoham, Rob Powers, and Trond Grenager. If multi-agent learning is the answer, what is the question? *Artificial intelligence*, 171(7):365–377, 2007. 3

55 Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008. 3, 12

56 Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797, 2019. 3, 10, 12

57 Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE transactions on cybernetics*, 50(9):3826–3839, 2020. 3

58 Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021. 3

59 Stefano V Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018. 3, 19, 26

60 Felipe Leno Da Silva and Anna Helena Reali Costa. A survey on transfer learning for multiagent reinforcement learning systems. *Journal of Artificial Intelligence Research*, 64:645–703, 2019. 3, 15, 36

61 Jim E Doran, SRJN Franklin, Nicholas R Jennings, and Timothy J Norman. On cooperation in multi-agent systems. *The Knowledge Engineering Review*, 12(3):309–314, 1997. 3

62 Liviu Panait and Sean Luke. Cooperative multi-agent learning: The state of the art. *Autonomous agents and multi-agent systems*, 11:387–434, 2005. 3

63 St John Grimbly, Jonathan Shock, and Arnu Pretorius. Causal multi-agent reinforcement learning: Review and open problems. *preprint arXiv:2111.06721*, 2021. 3, 30

64 Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Zixuan Ke, and Bing Liu. Open-world continual learning: Unifying novelty detection and continual learning. *preprint arXiv:2304.10038*, 2023. 3, 33

65 Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992. 5

66 Gerald Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995. 6

67 Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996. 6

68 Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. 6

69 Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas. Dueling network architectures for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 1995–2003, 2016. 6

70 Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2094–2100, 2016. 6

71 Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 449–458, 2017. 6

72 Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z. Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2017. 6

73 Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control. In *Proceedings of the International Conference on Machine Learning*, pages 2827–2836, 2017. 6

74 Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014, 1999. 6

75 Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pages 4565–4573, 2016. 7

76 David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin A. Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the International Conference on Machine Learning*, pages 387–395, 2014. 7

77 Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. 7

78 Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Rémi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. In *International Conference on Learning Representations*, 2017. 8

79 Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *preprint arXiv:1812.05905*, 2018. 8

80 Gerhard Weiss. *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. MIT press, 1999. 8

81 Drew Fudenberg and Jean Tirole. *Game theory*. MIT press, 1991. 8

82 Guillermo Owen. *Game theory*. Emerald Group Publishing, 2013. 8, 19

83 Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, 27(4):819–840, 2002. 12

84 Matthew Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable mdps. In *Proceedings of the 2015 AAAI Fall Symposia*, pages 29–37, 2015. 12

85 Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the international conference on machine learning*, pages 330–337, 1993. 13

86 Dean Foster, Dylan J Foster, Noah Golowich, and Alexander Rakhlin. On the complexity of multi-agent decision making: From learning in games to partial monitoring. In *the Annual Conference on Learning Theory*, pages 2678–2792, 2023. 14

87 Xueguang Lyu, Andrea Baisero, Yuchen Xiao, Brett Daley, and Christopher Amato. On centralized critics in multi-agent reinforcement learning. *Journal of Artificial Intelligence Research*, 77:295–354, 2023. 14

88 Yihe Zhou, Shunyu Liu, Yunpeng Qing, Kaixuan Chen, Tongya Zheng, Yanhao Huang, Jie Song, and Mingli Song. Is centralized training with decentralized execution framework centralized enough for marl? *preprint arXiv:2305.17352*, 2023. 14

89 Martin Lauer and Martin Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *Proceedings of the 17th International Conference on Machine Learning*, 2000. 16

90 Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference*, pages 746–752, 1998. 16, 31

91 Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2974–2982, 2018. 17, 19

92 Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 2961–2970, 2019. 19

93 Rihab Gorsane, Omayma Mahjoub, Ruan John de Kock, Roland Dubb, Siddarth Singh, and Arnu Pretorius. Towards a standardised performance evaluation protocol for cooperative marl. In *Advances in Neural Information Processing Systems*, pages 5510–5521, 2022. 19

94 Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. QPLEX: Duplex dueling multi-agent Q-learning. In *International Conference on Learning Representations*, 2021. 20, 23

95 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017. 20, 42

96 Yihan Wang, Beining Han, Tonghan Wang, Heng Dong, and Chongjie Zhang. Dop: Off-policy multi-agent decomposed policy gradients. In *International conference on learning representations*, 2020. 21, 22, 23

97 Jianyu Su, Stephen Adams, and Peter Beling. Value-decomposition multi-agent actor-critics. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11352–11360, 2021. 22

98 Xiaoteng Ma, Yiqin Yang, Chenghao Li, Yiwen Lu, Qianchuan Zhao, and Jun Yang. Modeling the interaction between agents in cooperative multi-agent reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 853–861, 2021. 22

99 Bei Peng, Tabish Rashid, Christian Schroeder de Witt, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. Facmac: Factored multi-agent centralised policy gradients. In *Advances in Neural Information Processing Systems*, pages 12208–12221, 2021. 22, 30, 31

100 Tianhao Zhang, Yueheng Li, Chen Wang, Guangming Xie, and Zongqing Lu. Fop: Factorizing optimal joint policy of maximum-entropy multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 12491–12500, 2021. 22

101 Jian Hu, Siyang Jiang, Seth Austin Harding, Haibin Wu, and Shih-wei Liao. Rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning. *preprint arXiv:2102.03479*, 2021. 22

102 Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2021. 23

103 Anuj Mahajan, Tabish Rashid, Mikayel Samvelyan, and Shimon Whiteson. MAVEN: multi-agent variational exploration. In *Advances in Neural Information Processing Systems*, pages 7611–7622, 2019. 23, 24

104 Tonghan Wang, Jianhao Wang, Yi Wu, and Chongjie Zhang. Influence-based multi-agent exploration. In *International Conference on Learning Representations*, 2019. 23, 24

105 Lulu Zheng, Jiarui Chen, Jianhao Wang, Jiamin He, Yujing Hu, Yingfeng Chen, Changjie Fan, Yang Gao, and Chongjie Zhang. Episodic multi-agent reinforcement learning with curiosity-driven exploration. In *Advances in Neural Information Processing Systems*, pages 3757–3769, 2021. 23, 24

106 Iou-Jen Liu, Unnat Jain, Raymond A Yeh, and Alexander Schwing. Cooperative exploration for multi-agent deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 6826–6836, 2021. 23, 24

107 Tarun Gupta, Anuj Mahajan, Bei Peng, Wendelin Böhmer, and Shimon Whiteson. Uneven: Universal value exploration for

multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 3930–3941, 2021. 23, 24

108 Shaowei Zhang, Jiahan Cao, Lei Yuan, Yang Yu, and De-Chuan Zhan. Self-motivated multi-agent exploration. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 476–484, 2023. 23, 24

109 Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2137–2145, 2016. 23, 25

110 Sai Qian Zhang, Qi Zhang, and Jieyu Lin. Efficient communication in multi-agent reinforcement learning via variance based control. In *Advances in Neural Information Processing Systems*, pages 3230–3239, 2019. 23, 25

111 Ziluo Ding, Tiejun Huang, and Zongqing Lu. Learning individually inferred communication for multi-agent cooperation. In *Advances in Neural Information Processing Systems*, pages 22069–22079, 2020. 23, 25

112 Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. Tarmac: Targeted multi-agent communication. In *Proceedings of the International Conference on Machine Learning*, pages 1538–1546, 2019. 23, 25

113 Lei Yuan, Jianhao Wang, Fuxiang Zhang, Chenghe Wang, Zongzhang Zhang, Yang Yu, and Chongjie Zhang. Multi-agent incentive communication via decentralized teammate modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9466–9474, 2022. 23, 26

114 Cong Guan, Feng Chen, Lei Yuan, Zongzhang Zhang, and Yang Yu. Efficient communication via self-supervised information aggregation for online and offline multi-agent reinforcement learning. *preprint arXiv:2302.09605*, 2023. 23, 25, 31

115 Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *Proceedings of the International conference on machine learning*, pages 4218–4227, 2018. 23, 26

116 Georgios Papoudakis and Stefano V Albrecht. Variational autoencoders for opponent modeling in multi-agent systems. *preprint arXiv:2001.10829*, 2020. 23, 26

117 Georgios Papoudakis, Filippos Christianos, and Stefano Albrecht. Agent modelling under partial observability for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 19210–19222, 2021. 23, 26

118 Annie Xie, Dylan Losey, Ryan Tolsma, Chelsea Finn, and Dorsa Sadigh. Learning latent representations to influence multi-agent interaction. In *Conference on robot learning*, pages 575–588, 2021. 23, 26

119 Xiaopeng Yu, Jiechuan Jiang, Wanpeng Zhang, Haobin Jiang, and Zongqing Lu. Model-based opponent modeling. In *Advances in Neural Information Processing Systems*, pages 28208–28221, 2022. 23, 26, 28

120 Lei Yuan, Chenghe Wang, Jianhao Wang, Fuxiang Zhang, Feng Chen, Cong Guan, Zongzhang Zhang, Chongjie Zhang, and Yang Yu. Multi-agent concentrative coordination with decentralized task representation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 599–605, 2022. 23, 26

121 Jiaming Song, Hongyu Ren, Dorsa Sadigh, and Stefano Ermon. Multi-agent generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pages 7472–7483, 2018. 23, 27

122 Lantao Yu, Jiaming Song, and Stefano Ermon. Multi-agent adversarial inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 7194–7201, 2019. 23, 27

123 Minghuan Liu, Ming Zhou, Weinan Zhang, Yuzheng Zhuang, Jun Wang, Wulong Liu, and Yong Yu. Multi-agent interactions modeling with correlated policies. In *International Conference on Learning Representations*, 2019. 23, 27

124 Caroline Wang, Ishan Durugkar, Elad Liebman, and Peter Stone. Distributed multi-agent reinforcement learning for distribution matching. *preprint arXiv:2206.00233*, 2022. 23, 27

125 Daniël Willemsen, Mario Coppola, and Guido CHE de Croon. Mambpo: Sample-efficient multi-robot reinforcement learning using learned world models. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5635–5640, 2021. 23, 28

126 Weinan Zhang, Xihuai Wang, Jian Shen, and Ming Zhou. Model-based multi-agent policy optimization with adaptive opponent-wise rollouts. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3384–3391, 2021. 23, 28

127 Zhiwei Xu, Bin Zhang, Yuan Zhan, Yunpeng Baiia, Guoliang Fan, et al. Mingling foresight with imagination: Model-based cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, pages 11327–11340, 2022. 23, 28

128 Vladimir Egorov and Alexei Shpilman. Scalable multi-agent model-based reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 381–390, 2022. 23, 28

129 Zhizun Wang and David Meger. Leveraging world model disentanglement in value-based multi-agent reinforcement learning. *preprint arXiv:2309.04615*, 2023. 23, 28

130 S Ahilan and P Dayan. Feudal multi-agent hierarchies for cooperative reinforcement learning. In *Workshop on Structure & Priors in Reinforcement Learning (SPiRL 2019) at ICLR 2019*, pages 1–11, 2019. 24, 28

131 Jiachen Yang, Igor Borovikov, and Hongyuan Zha. Hierarchical cooperative multi-agent reinforcement learning with skill discovery. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 1566–1574, 2020. 24, 28

132 Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, and Chongjie Zhang. Rode: Learning roles to decompose multi-agent tasks. In *International Conference on Learning Representations*, 2020. 24, 28

133 Shariq Iqbal, Robby Costales, and Fei Sha. Alma: Hierarchical learning for composite multi-agent tasks. In *Advances in Neural Information Processing Systems*, pages 7155–7166, 2022. 24, 28

134 Zhiwei Xu, Yunpeng Bai, Bin Zhang, Dapeng Li, and Guoliang Fan. Haven: hierarchical cooperative multi-agent reinforcement learning with dual coordination mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11735–11743, 2023. 24, 28

135 Carlos Guestrin, Michail G. Lagoudakis, and Ronald Parr. Coordinated reinforcement learning. In *Proceedings of the International Conference Machine Learning*, pages 227–234, 2002. 24, 29

136 Wendelin Boehmer, Vitaly Kurin, and Shimon Whiteson. Deep coordination graphs. In *Proceedings of the International Conference on Machine Learning*, pages 980–991, 2020. 24, 29

137 Sheng Li, Jayesh K Gupta, Peter Morales, Ross Allen, and Mykel J Kochenderfer. Deep implicit coordination graphs for multi-agent reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 764–772, 2021. 24, 29

138 Yaru Niu, Rohan Paleja, and Matthew Gombolay. Multi-agent graph-attention communication and teaming. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 964–973, 2021. 24, 29

139 Jiechuan Jiang and Zongqing Lu. Learning attentional communication for multi-agent cooperation. In *Advances in Neural Information Processing Systems*, pages 7265–7275, 2018. 24, 29

140 Tonghan Wang, Liang Zeng, Weijun Dong, Qianlan Yang, Yang Yu, and Chongjie Zhang. Context-aware sparse deep coordination graphs. In *International Conference on Learning Representations*, 2022. 24, 29, 31

141 Zichuan Liu, Yuanyang Zhu, and Chunlin Chen. Na2q: Neural attention additive model for interpretable multi-agent q-learning. *preprint arXiv:2304.13383*, 2023. 24, 30

142 Chuming Li, Jie Liu, Yinmin Zhang, Yuhong Wei, Yazhe Niu, Yaodong Yang, Yu Liu, and Wanli Ouyang. Ace: cooperative multi-agent q-learning with bidirectional action-dependency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8536–8544, 2023. 24, 30

143 Jiachen Yang, Alireza Nakhaei, David Isele, Kikuo Fujimura, and Hongyuan Zha. Cm3: Cooperative multi-goal multi-stage multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2019. 24, 30

144 Haotian Fu, Hongyao Tang, Jianye Hao, Zihan Lei, Yingfeng Chen, and Changjie Fan. Deep multi-agent reinforcement learning with discrete-continuous hybrid action spaces. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2329–2335, 2019. 24, 30

145 Jiayu Chen, Yuanxin Zhang, Yuanfan Xu, Huimin Ma, Huazhong Yang, Jiaming Song, Yu Wang, and Yi Wu. Variational automatic curriculum learning for sparse-reward cooperative multi-agent problems. In *Advances in Neural Information Processing Systems*, pages 9681–9693, 2021. 24, 30

146 Felipe Leno Da Silva, Garrett Warnell, Anna Helena Reali Costa, and Peter Stone. Agents teaching agents: a survey on inter-agent transfer learning. *Autonomous Agents and Multi-Agent Systems*, 34:1–17, 2020. 24, 30

147 Niko A Grupen, Bart Selman, and Daniel D Lee. Cooperative multi-agent fairness and equivariant policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9350–9359, 2022. 24, 30

148 Felipe Leno Da Silva, Ruben Glatt, and Anna Helena Reali Costa. Moo-mdp: An object-oriented representation for cooperative multiagent reinforcement learning. *IEEE Transactions on Cybernetics*, 49(2):567–579, 2017. 24, 30

149 Jianye Hao, Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Zhaopeng Meng, Peng Liu, and Zhen Wang. Exploration in deep reinforcement learning: From single-agent to multiagent domain. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 22

150 Woojun Kim and Youngchul Sung. An adaptive entropy-regularization framework for multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 16829–16852, 2023. 24

151 Shariq Iqbal and Fei Sha. Coordinated exploration via intrinsic rewards for multi-agent reinforcement learning. *preprint arXiv:1905.12127*, 2019. 24

152 Alberto Viseras, Thomas Wiedemann, Christoph Manss, Lukas Magel, Joachim Mueller, Dmitriy Shutin, and Luis Merino.

Decentralized multi-agent exploration with online-learning of gaussian processes. In *2016 IEEE international conference on robotics and automation*, pages 4222–4229, 2016. 24

153 Hans J He, Alec Koppel, Amrit Singh Bedi, Daniel J Stilwell, Mazen Farhood, and Benjamin Biggs. Decentralized multi-agent exploration with limited inter-agent communications. In *2023 IEEE International Conference on Robotics and Automation*, pages 5530–5536, 2023. 24

154 M Baglietto, M Paolucci, L Scardovi, and R Zoppoli. Information-based multi-agent exploration. In *Proceedings of the Third International Workshop on Robot Motion and Control*, pages 173–179, 2002. 24

155 Jingtian Yan, Xingqiao Lin, Zhongqiang Ren, Shiqi Zhao, Jieqiong Yu, Chao Cao, Peng Yin, Ji Zhang, and Sebastian Scherer. Mui-tare: Cooperative multi-agent exploration with unknown initial position. *IEEE Robotics and Automation Letters*, 2023. 24

156 Mithun Chakraborty, Kai Yee Phoebe Chua, Sanmay Das, and Brendan Juba. Coordinated versus decentralized exploration in multi-agent multi-armed bandits. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 164–170, 2017. 24

157 Stefanos Leonardos, Georgios Piliouras, and Kelly Spendlove. Exploration-exploitation in multi-agent competition: Convergence with bounded rationality. In *Advances in Neural Information Processing Systems*, pages 26318–26331, 2021. 24

158 Yonghyeon Jo, Sunwoo Lee, Junghyuk Yum, and Seungyul Han. Fox: Formation-aware exploration in multi-agent reinforcement learning. *preprint arXiv:2308.11272*, 2023. 24

159 Jihwan Oh, Joonkee Kim, Minchan Jeong, and Se-Young Yun. Toward risk-based optimistic exploration for cooperative multi-agent reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 1597–1605, 2023. 24, 34, 40

160 Pei Xu, Junge Zhang, and Kaiqi Huang. Exploration via joint policy diversity for sparse-reward multi-agent tasks. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 326–334, 2023. 24

161 Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems*, pages 2244–2252, 2016. 25, 31

162 Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. Learning when to communicate at scale in multiagent cooperative and competitive tasks. In *International Conference on Learning Representations*, 2019. 25

163 Hangyu Mao, Zhengchao Zhang, Zhen Xiao, Zhibo Gong, and Yan Ni. Learning agent communication under limited bandwidth by message pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5142–5149, 2020. 25

164 Di Xue, Lei Yuan, Zongzhang Zhang, and Yang Yu. Efficient multi-agent communication via shapley message value. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 578–584, 2022. 25

165 Hangyu Mao, Zhengchao Zhang, Zhen Xiao, Zhibo Gong, and Yan Ni. Learning multi-agent communication with double attentional deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 34(1):1–34, 2020. 25

166 Tonghan Wang, Jianhao Wang, Chongyi Zheng, and Chongjie Zhang. Learning nearly decomposable value functions via communication minimization. In *International Conference on Learning Representations*, 2020. 25, 31

167 Sai Qian Zhang, Qi Zhang, and Jieyu Lin. Succinct and robust multi-agent communication with temporal message control. In *Advances in Neural Information Processing Systems*, pages 17271–17282, 2020. 25

168 Lei Yuan, Tao Jiang, Lihe Li, Feng Chen, Zongzhang Zhang, and Yang Yu. Robust multi-agent communication via multi-view message certification. *preprint arXiv:2305.13936*, 2023. 26, 34, 39

169 Yanchao Sun, Ruijie Zheng, Parisa Hassanzadeh, Yongyuan Liang, Soheil Feizi, Sumitra Ganesh, and Furong Huang. Certifiably robust policy learning against adversarial multi-agent communication. In *The International Conference on Learning Representations*, 2022. 26

170 Lei Yuan, Feng Chen, Zhongzhang Zhang, and Yang Yu. Communication-robust multi-agent learning by adaptable auxiliary multi-agent adversary generation. *preprint arXiv:2305.05116*, 2023. 26, 34, 39

171 Jing-Cheng Pang, Pengyuan Wang, Kaiyuan Li, Xiong-Hui Chen, Jiacheng Xu, Zongzhang Zhang, and Yang Yu. Language model self-improvement by reinforcement learning contemplation. *preprint arXiv:2305.14483*, 2023. 26

172 Yuanfei Wang, Jing Xu, Yizhou Wang, et al. Tom2c: Target-oriented multi-agent communication and cooperation with theory of mind. In *International Conference on Learning Representations*, 2021. 26

173 Michael Shum, Max Kleiman-Weiner, Michael L Littman, and Joshua B Tenenbaum. Theory of minds: Understanding behavior in groups through inverse planning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6163–6170, 2019. 26

174 Emre Erdogan, Frank Dignum, Rineke Verbrugge, and Pinar Yolum. Abstracting minds: Computational theory of mind for human-agent collaboration. In *Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence, Amsterdam,*, pages 199–211, 2022. 26

175 Christelle Langley, Bogdan Ionut Cirstea, Fabio Cuzzolin, and Barbara J Sahakian. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and ai: A review. *Frontiers in Artificial Intelligence*, 5:62, 2022. 26

176 Jiahan Cao, Lei Yuan, Jianhao Wang, Shaowei Zhang, Chongjie Zhang, Yang Yu, and De-Chuan Zhan. Linda: Multi-agent local information decomposition for awareness of teammates. *Science China Information Sciences*, 66(8):182101, 2023. 26

177 Woodrow Zhouyuan Wang, Andy Shih, Annie Xie, and Dorsa Sadigh. Influencing towards stable multi-agent interactions. In *Conference on robot learning*, pages 1132–1143, 2022. 26

178 Ran Tian, Masayoshi Tomizuka, Anca D Dragan, and Andrea Bajcsy. Towards modeling and influencing the dynamics of human learning. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 350–358, 2023. 26

179 Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. Modeling others using oneself in multi-agent reinforcement learning. In *Proceedings of the International conference on machine learning*, pages 4257–4266, 2018. 26

180 Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. Agent modeling as auxiliary task for deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence and interactive digital entertainment*, pages 31–37, 2019. 26

181 Zhang-Wei Hong, Shih-Yang Su, Tzu-Yun Shann, Yi-Hsiang Chang, and Chun-Yi Lee. A deep policy inference q-network for multi-agent systems. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 1388–1396, 2018. 26

182 Hangyu Mao, Zhengchao Zhang, Zhen Xiao, and Zhibo Gong. Modelling the dynamic joint policy of teammates with attention multi-agent ddpg. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 1108–1116, 2019. 26

183 Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *The International Conference on Learning Representations*, 2016. 26

184 Ying Wen, Yaodong Yang, Rui Luo, Jun Wang, and Wei Pan. Probabilistic recursive reasoning for multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2018. 26

185 Julien Roy, Paul Barde, Félix Harvey, Derek Nowrouzezahrai, and Chris Pal. Promoting coordination through policy regularization in multi-agent deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 15774–15785, 2020. 26

186 Maryam Zare, Parham M Kebria, Abbas Khosravi, and Saeid Nahavandi. A survey of imitation learning: Algorithms, recent developments, and challenges. *preprint arXiv:2309.02473*, 2023. 27

187 Hoang M Le, Yisong Yue, Peter Carr, and Patrick Lucey. Coordinated multi-agent imitation learning. In *Proceedings of the International Conference on Machine Learning*, pages 1995–2003, 2017. 27

188 Eric Zhan, Stephan Zheng, Yisong Yue, Long Sha, and Patrick Lucey. Generating multi-agent trajectories using programmatic weak supervision. In *International Conference on Learning Representations*, 2018. 27

189 Hongwei Wang, Lantao Yu, Zhangjie Cao, and Stefano Ermon. Multi-agent imitation learning with copulas. In *Machine Learning and Knowledge Discovery in Databases*, pages 139–156, 2021. 27

190 Roger B Nelsen. *An introduction to copulas*. Springer, 2006. 27

191 Nate Gruver, Jiaming Song, Mykel J Kochenderfer, and Stefano Ermon. Multi-agent adversarial inverse reinforcement learning with latent variables. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 1855–1857, 2020. 27

192 The Viet Bui, Tien Mai, and Thanh Hong Nguyen. Inverse factorized q-learning for cooperative multi-agent imitation learning. *preprint arXiv:2310.06801*, 2023. 27

193 Raunak P Bhattacharyya, Derek J Phillips, Blake Wulfe, Jeremy Morton, Alex Kuefler, and Mykel J Kochenderfer. Multi-agent imitation learning for driving simulation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1534–1539, 2018. 27

194 Justin Fu, Andrea Tacchetti, Julien Perolat, and Yoram Bachrach. Evaluating strategic structures in multi-agent inverse reinforcement learning. *Journal of Artificial Intelligence Research*, 71:925–951, 2021. 27

195 Xingyu Wang and Diego Klabjan. Competitive multi-agent inverse reinforcement learning with sub-optimal demonstrations. In *Proceedings of the International Conference on Machine Learning*, pages 5143–5151, 2018. 27

196 Yang Chen, Libo Zhang, Jiamou Liu, and Michael Witbrock. Adversarial inverse reinforcement learning for mean field games. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 1088–1096, 2023. 27

197 Shicheng Liu and Minghui Zhu. Distributed inverse constrained reinforcement learning for multi-agent systems. In *Advances in Neural Information Processing Systems*, pages 33444–33456, 2022. 27

198 Xin Zhang, Weixiao Huang, Yanhua Li, Renjie Liao, and Ziming Zhang. Imitation learning from inconcurrent multi-agent interactions. In *Proceedings of the IEEE Annual Conference on Decision and Control*, pages 43–48, 2021. 27

199 Fan-Ming Luo, Tian Xu, Hang Lai, Xiong-Hui Chen, Weinan Zhang, and Yang Yu. A survey on model-based reinforcement learning. *preprint arXiv:2206.09328*, 2022. 28

200 Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1):1–118, 2023. 28

201 Michael Janner, Justin Fu, Marvin Zhang, , and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, pages 12498–12509, 2019. 28

202 Young Joon Park, Yoon Sang Cho, and Seoung Bum Kim. Multi-agent reinforcement learning with approximate model learning for competitive games. *PloS one*, 14(9):e0222215, 2019. 28

203 Qizhen Zhang, Chris Lu, Animesh Garg, and Jakob Foerster. Centralized model and exploration policy for multi-agent rl. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 1500–1508, 2022. 28

204 Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2019. 28

205 Rose Wang, J Chase Kew, Dennis Lee, Tsang-Wei Lee, Tingnan Zhang, Brian Ichter, Jie Tan, and Aleksandra Faust. Model-based reinforcement learning for decentralized multiagent rendezvous. In *Conference on Robot Learning*, pages 711–725, 2021. 28

206 Wenli Xiao, Yiwei Lyu, and John Dolan. Model-based dynamic shielding for safe and efficient multi-agent reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 1587–1596, 2023. 28, 34, 40

207 Anuj Mahajan, Mikayel Samvelyan, Lei Mao, Viktor Makoviychuk, Animesh Garg, Jean Kossaifi, Shimon Whiteson, Yuke Zhu, and Animashree Anandkumar. Tesseract: Tensorised actors for multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 7301–7312, 2021. 28

208 Yali Du, Chengdong Ma, Yuchen Liu, Runji Lin, Hao Dong, Jun Wang, and Yaodong Yang. Scalable model-based policy optimization for decentralized networked systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 9019–9026, 2022. 28

209 Woojun Kim, Jongeui Park, and Youngchul Sung. Communication in multi-agent reinforcement learning: Intention sharing. In *International Conference on Learning Representations*, 2020. 28

210 Ziluo Ding, Kefan Su, Weixin Hong, Liwen Zhu, Tiejun Huang, and Zongqing Lu. Multi-agent sequential decision-making via communication. *preprint arXiv:2209.12713*, 2022. 28

211 Shuai Han, Mehdi Dastani, and Shihan Wang. Model-based sparse communication in multi-agent reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 439–447, 2023. 28

212 Barna Pásztor, Andreas Krause, and Ilija Bogunovic. Efficient model-based multi-agent mean-field reinforcement learning. *Transactions on Machine Learning Research*, 2023. 28

213 Pier Giuseppe Sessa, Maryam Kamgarpour, and Andreas Krause. Efficient model-based multi-agent reinforcement learning via optimistic equilibrium computation. In *Proceedings of the International Conference on Machine Learning*, pages 19580–19597, 2022. 28

214 Paul Barde, Jakob Foerster, Derek Nowrouzezahrai, and Amy Zhang. A model-based solution to the offline multi-agent reinforcement learning coordination problem. *preprint arXiv:2305.17198*, 2023. 28

215 Dongge Han, Chris Xiaoxuan Lu, Tomasz Michalak, and Michael Wooldridge. Multiagent model-based credit assignment for continuous control. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 571–579, 2022. 28

216 Chao Qian, Yang Yu, and Zhi-Hua Zhou. Subset selection by pareto optimization. In *Advances in Neural Information Processing Systems*, pages 1774–1782, 2015. 28, 38

217 Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(5):1–35, 2021. 28

218 Rajbala Makar, Sridhar Mahadevan, and Mohammad Ghavamzadeh. Hierarchical multi-agent reinforcement learning. In

Proceedings of the international conference on Autonomous agents, pages 246–253, 2001. 28

219 Hongyao Tang, Jianye Hao, Tangjie Lv, Yingfeng Chen, Zongzhang Zhang, Hangtian Jia, Chunxu Ren, Yan Zheng, Zhaopeng Meng, Changjie Fan, et al. Hierarchical deep multiagent reinforcement learning with temporal abstraction. *preprint arXiv:1809.09332*, 2018. 28

220 Thomy Phan, Fabian Ritz, Lenz Belzner, Philipp Altmann, Thomas Gabor, and Claudia Linnhoff-Popien. Vast: Value function factorization with variable agent sub-teams. In *Advances in Neural Information Processing Systems*, pages 24018–24032, 2021. 28

221 Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*, 2018. 28

222 Holger Friedrich, Oliver Rogalla, and Ru¨ diger Dillmann. Integrating skills into multi-agent systems. *Journal of Intelligent Manufacturing*, 9:119–127, 1998. 29

223 Yuntao Liu, Yuan Li, Xinhai Xu, Yong Dou, and Donghong Liu. Heterogeneous skill learning for multi-agent tasks. In *Advances in Neural Information Processing Systems*, pages 37011–37023, 2022. 29

224 Shuncheng He, Jianzhun Shao, and Xiangyang Ji. Skill discovery of coordination in multi-agent reinforcement learning. *preprint arXiv:2006.04021*, 2020. 29

225 Rundong Wang, Longtao Zheng, Wei Qiu, Bowei He, Bo An, Zinovi Rabinovich, Yujing Hu, Yingfeng Chen, Tangjie Lv, and Changjie Fan. Towards skilled population curriculum for multi-agent reinforcement learning. *preprint arXiv:2302.03429*, 2023. 29

226 Jiayu Chen, Jingdi Chen, Tian Lan, and Vaneet Aggarwal. Scalable multi-agent covering option discovery based on kronecker graphs. In *Advances in Neural Information Processing Systems*, pages 30406–30418, 2022. 29

227 Jing-Cheng Pang, Xin-Yu Yang, Si-Hang Yang, and Yang Yu. Natural language-conditioned reinforcement learning with inside-out task language development and translation. *preprint arXiv:2302.09368*, 2023. 29

228 Carlos Guestrin, Shobha Venkataraman, and Daphne Koller. Context-specific multiagent coordination and planning with factored mdps. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence*, pages 253–259, 2002. 29

229 Shanjun Cheng. *Coordinating decentralized learning and conflict resolution across agent boundaries*. PhD thesis, The University of North Carolina at Charlotte, 2012. 29

230 Jelle R Kok and Nikos Vlassis. Collaborative multiagent reinforcement learning by payoff propagation. *Journal of Machine Learning Research*, 7:1789–1828, 2006. 29

231 Chongjie Zhang and Victor Lesser. Coordinating multi-agent reinforcement learning with limited communication. In *Proceedings of the international conference on Autonomous agents and multi-agent systems*, pages 1101–1108, 2013. 29, 35

232 Jacopo Castellini, Frans A Oliehoek, Rahul Savani, and Shimon Whiteson. The representational capacity of action-value networks for multi-agent reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 1862–1864, 2019. 29

233 Yipeng Kang, Tonghan Wang, Qianlan Yang, Xiaoran Wu, and Chongjie Zhang. Non-linear coordination graphs. In *Advances in Neural Information Processing Systems*, pages 25655–25666, 2022. 29

234 Qianlan Yang, Weijun Dong, Zhizhou Ren, Jianhao Wang, Tonghan Wang, and Chongjie Zhang. Self-organized polynomial-time coordination graphs. In *Proceedings of the International Conference on Machine Learning*, pages 24963–24979, 2022. 29

235 Junjie Sheng, Xiangfeng Wang, Bo Jin, Wenhao Li, Jun Wang, Junchi Yan, Tsung-Hui Chang, and Hongyuan Zha. Learning structured communication for multi-agent reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 436–438, 2023. 29

236 Jakub Grudzien Kuba, Muning Wen, Linghui Meng, Haifeng Zhang, David Mguni, Jun Wang, Yaodong Yang, et al. Settling the variance of multi-agent policy gradients. In *Advances in Neural Information Processing Systems*, pages 13458–13470, 2021. 30

237 Xihuai Wang, Zheng Tian, Ziyu Wan, Ying Wen, Jun Wang, and Weinan Zhang. Order matters: Agent-by-agent policy optimization. In *The International Conference on Learning Representations*, 2023. 30

238 Zehao Dou, Jakub Grudzien Kuba, and Yaodong Yang. Understanding value decomposition algorithms in deep cooperative multi-agent reinforcement learning. *preprint arXiv:2202.04868*, 2022. 30

239 Jianghai Hu. *Multi-agent coordination: Theory and applications*. University of California, Berkeley, 2003. 30

240 Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the Conference on Autonomous Agents and MultiAgent Systems*, pages

464–473, 2017. 30, 31

241 Yuchen Xiao, Weihao Tan, and Christopher Amato. Asynchronous actor-critic for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 35:4385–4400, 2022. 30

242 Hancheng Zhang, Guozheng Li, Chi Harold Liu, Guoren Wang, and Jian Tang. Himacmic: Hierarchical multi-agent deep reinforcement learning with dynamic asynchronous macro strategy. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3239–3248, 2023. 30

243 Qingxu Fu, Tenghai Qiu, Jianqiang Yi, Zhiqiang Pu, and Shiguang Wu. Concentration network for reinforcement learning of large-scale multi-agent systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9341–9349, 2022. 30, 31

244 Wei Qiu, Weixun Wang, Rundong Wang, Bo An, Yujing Hu, Svetlana Obraztsova, Zinovi Rabinovich, Jianye Hao, Yingfeng Chen, and Changjie Fan. Off-beat multi-agent reinforcement learning. *preprint arXiv:2205.13718*, 2022. 30

245 Benjamin Ellis, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob N Foerster, and Shimon Whiteson. Smacv2: An improved benchmark for cooperative multi-agent reinforcement learning. *preprint arXiv:2212.07489*, 2022. 30

246 Adam Michalski, Filippos Christianos, and Stefano V Albrecht. Smaclite: A lightweight environment for multi-agent reinforcement learning. *preprint arXiv:2305.05566*, 2023. 30

247 Karol Kurach, Anton Raichuk, Piotr Stanczyk, Michal Zajac, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4501–4510, 2020. 30, 31

248 Lianmin Zheng, Jiacheng Yang, Han Cai, Ming Zhou, Weinan Zhang, Jun Wang, and Yong Yu. Magent: A many-agent reinforcement learning platform for artificial collective intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 30, 31

249 Claude Formanek, Asad Jeewa, Jonathan Shock, and Arnu Pretorius. Off-the-grid marl: Datasets and baselines for offline multi-agent reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, page 2442–2444, 2023. 30, 31, 36

250 Qirui Mi, Siyu Xia, Yan Song, Haifeng Zhang, Shenghao Zhu, and Jun Wang. Taxai: A dynamic economic simulator and benchmark for multi-agent reinforcement learning. *preprint arXiv:2309.16307*, 2023. 30

251 Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*, 2021. 30, 31

252 J Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. Pettingzoo: Gym for multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 15032–15043, 2021. 30

253 Siyi Hu, Yifan Zhong, Minquan Gao, Weixun Wang, Hao Dong, Zhihui Li, Xiaodan Liang, Xiaojun Chang, and Yaodong Yang. Marllib: Extending rllib for multi-agent reinforcement learning. *preprint arXiv:2210.13708*, 2022. 30

254 Ming Zhang, Shenghan Zhang, Zhenjie Yang, Lekai Chen, Jinliang Zheng, Chao Yang, Chuming Li, Hang Zhou, Yazhe Niu, and Yu Liu. Gobigger: A scalable platform for cooperative-competitive multi-agent interactive simulation. In *International Conference on Learning Representations*, 2022. 31

255 Diego Perez-Liebana, Katja Hofmann, Sharada Prasanna Mohanty, Noburu Kuno, Andre Kramer, Sam Devlin, Raluca D Gaina, and Daniel Ionita. The multi-agent reinforcement learning in malmö (marlö) competition. *preprint arXiv:1901.08129*, 2019. 31

256 Cinjon Resnick, Wes Eldridge, David Ha, Denny Britz, Jakob Foerster, Julian Togelius, Kyunghyun Cho, and Joan Bruna. Pommerman: A multi-agent playground. *preprint arXiv:1809.07124*, 2018. 31

257 Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020. 31

258 Micah Carroll, Rohin Shah, Mark K. Ho, Tom Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca D. Dragan. On the utility of learning about humans for human-ai coordination. In *Advances in Neural Information Processing Systems*, pages 5175–5186, 2019. 31, 42

259 Joseph Suarez, Yilun Du, Clare Zhu, Igor Mordatch, and Phillip Isola. The neural mmo platform for massively multiagent research. *preprint arXiv:2110.07594*, 2021. 31

260 Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. In *International Conference on Learning Representations*, 2019. 30, 31

261 Hangtian Jia, Yujing Hu, Yingfeng Chen, Chunxu Ren, Tangjie Lv, Changjie Fan, and Chongjie Zhang. Fever basketball: A complex, flexible, and asynchronized sports game environment for multi-agent reinforcement learning. *preprint arXiv:2012.03204*, 2020. 30, 31

262 Daniel Krajzewicz. Traffic simulation with sumo–simulation of urban mobility. *Fundamentals of traffic simulation*, pages 269–293, 2010. 31

263 Huichu Zhang, Siyuan Feng, Chang Liu, Yaoyao Ding, Yichen Zhu, Zihan Zhou, Weinan Zhang, Yong Yu, Haiming Jin, and Zhenhui Li. Cityflow: A multi-agent reinforcement learning environment for large scale city traffic scenario. In *The World Wide Web Conference,*, pages 3620–3624, 2019. 31

264 Roni Stern, Nathan R. Sturtevant, Ariel Felner, Sven Koenig, Hang Ma, Thayne T. Walker, Jiaoyang Li, Dor Atzmon, Liron Cohen, T. K. Satish Kumar, Roman Barták, and Eli Boyarski. Multi-agent pathfinding: Definitions, variants, and benchmarks. In *Proceedings of the Twelfth International Symposium on Combinatorial Search*, pages 151–159, 2019. 31

265 Sharada Mohanty, Erik Nygren, Florian Laurent, Manuel Schneider, Christian Scheller, Nilabha Bhattacharya, Jeremy Watson, Adrian Egli, Christian Eichenberger, Christian Baumberger, et al. Flatland-rl: Multi-agent reinforcement learning on trains. *preprint arXiv:2012.05893*, 2020. 31

266 Ming Zhou, Jun Luo, Julian Villella, Yaodong Yang, David Rusu, Jiayu Miao, Weinan Zhang, Montgomery Alban, Iman Fadakar, Zheng Chen, et al. Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving. *preprint arXiv:2010.09776*, 2020. 31

267 Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3461–3475, 2022. 31

268 Xuehai Pan, Mickel Liu, Fangwei Zhong, Yaodong Yang, Song-Chun Zhu, and Yizhou Wang. Mate: Benchmarking multi-agent reinforcement learning in distributed target coverage control. In *Advances in Neural Information Processing Systems*, pages 27862–27879, 2022. 31

269 Reza Torbati, Shubbham Lohiya, Shivika Singh, Meher S Nigam, and Harish Ravichandar. Marbler: An open platform for standarized evaluation of multi-robot reinforcement learning algorithms. *preprint arXiv:2307.03891*, 2023. 31

270 Xianliang Yang, Zhihao Liu, Wei Jiang, Chuheng Zhang, Li Zhao, Lei Song, and Jiang Bian. A versatile multi-agent reinforcement learning benchmark for inventory management. *preprint arXiv:2306.07542*, 2023. 31

271 Xiaoteng Ma QihanLiu, Yuhua Jiang. Light aircraft game: A lightweight, scalable, gym-wrapped aircraft competitive environment with baseline reinforcement learning algorithms. https://github.com/liuqh16/CloseAirCombat, 2022. 31

272 Fang Gao, Si Chen, Mingqiang Li, and Bincheng Huang. Maca: a multi-agent reinforcement learning platform for collective intelligence. In *2019 IEEE 10th International Conference on Software Engineering and Service Science*, pages 108–111, 2019. 31

273 Yuyu Yuan, Pengqian Zhao, Ting Guo, and Hongpu Jiang. Counterfactual-based action evaluation algorithm in multi-agent reinforcement learning. *Applied Sciences*, 12(7), 2022. 31

274 Shangding Gu, Jakub Grudzien Kuba, Yuanpei Chen, Yali Du, Long Yang, Alois Knoll, and Yaodong Yang. Safe multi-agent reinforcement learning for multi-robot control. *Artificial Intelligence*, 319:103905, 2023. 31, 34, 40

275 F Bahrpeyma and D Reichelt. A review of the applications of multi-agent reinforcement learning in smart factories. *Frontiers in Robotics and AI*, 9:1027340–1027340, 2022. 30

276 Lorenzo Canese, Gian Carlo Cardarilli, Luca Di Nunzio, Rocco Fazzolari, Daniele Giardino, Marco Re, and Sergio Spanò. Multi-agent reinforcement learning: A review of challenges and applications. *Applied Sciences*, 11(11):4948, 2021. 30

277 Tianxu Li, Kun Zhu, Nguyen Cong Luong, Dusit Niyato, Qihui Wu, Yang Zhang, and Bing Chen. Applications of multi-agent reinforcement learning in future internet: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 24(2):1240–1279, 2022. 30

278 Ziyuan Zhou, Guanjun Liu, and Ying Tang. Multi-agent reinforcement learning: Methods, applications, visionary prospects, and challenges. *preprint arXiv:2305.10091*, 2023. 30

279 Zun Li, Marc Lanctot, Kevin R McKee, Luke Marris, Ian Gemp, Daniel Hennes, Paul Muller, Kate Larson, Yoram Bachrach, and Michael P Wellman. Combining tree-search, generative models, and nash bargaining concepts in game-theoretic reinforcement learning. *preprint arXiv:2302.00797*, 2023. 30

280 Zhijian Zhang, Haozheng Li, Luo Zhang, Tianyin Zheng, Ting Zhang, Xiong Hao, Xiaoxin Chen, Min Chen, Fangxu Xiao, and Wei Zhou. Hierarchical reinforcement learning for multi-agent moba game. *preprint arXiv:1901.08004*, 2019. 30

281 Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T Connor, Neil Burch, Thomas Anthony, et al. Mastering the game of stratego with model-free multiagent reinforcement

learning. *Science*, 378(6623):990–996, 2022. 30

282 Yang Li, Kun Xiong, Yingping Zhang, Jiangcheng Zhu, Stephen Mcaleer, Wei Pan, Jun Wang, Zonghong Dai, and Yaodong Yang. Jiangjun: Mastering xiangqi by tackling non-transitivity in two-player zero-sum games. *preprint arXiv:2308.04719*, 2023. 30

283 Xiangyu Zhao and Sean B Holden. Towards a competitive 3-player mahjong ai using deep reinforcement learning. In *2022 IEEE Conference on Games*, pages 524–527, 2022. 30

284 Daochen Zha, Jingru Xie, Wenye Ma, Sheng Zhang, Xiangru Lian, Xia Hu, and Ji Liu. Douzero: Mastering doudizhu with self-play deep reinforcement learning. In *international conference on machine learning*, pages 12333–12344, 2021. 30

285 Raymond A Yeh, Alexander G Schwing, Jonathan Huang, and Kevin Murphy. Diverse generation for multi-agent sports games. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4610–4619, 2019. 30

286 Dimitrios Troullinos, Georgios Chalkiadakis, Ioannis Papamichail, and Markos Papageorgiou. Collaborative multiagent decision making for lane-free autonomous driving. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 1335–1343, 2021. 30

287 Tong Wang, Jiahua Cao, and Azhar Hussain. Adaptive traffic signal control for large-scale scenario with cooperative group-based multi-agent reinforcement learning. *Transportation research part C: emerging technologies*, 125:103046, 2021. 32

288 Siyuan Chen, Meiling Wang, Wenjie Song, Yi Yang, and Mengyin Fu. Multi-agent reinforcement learning-based twin-vehicle fair cooperative driving in dynamic highway scenarios. In *2022 IEEE International Conference on Intelligent Transportation Systems*, pages 730–736, 2022. 32

289 Sangwoo Jeon, Hoeun Lee, Vishnu Kumar Kaliappan, Tuan Anh Nguyen, Hyungeun Jo, Hyeonseo Cho, and Dugki Min. Multiagent reinforcement learning based on fusion-multiactor-attention-critic for multiple-unmanned-aerial-vehicle navigation control. *Energies*, 15(19):7426, 2022. 32

290 Shutong Chen, Guanjun Liu, Ziyuan Zhou, Kaiwen Zhang, and Jiacun Wang. Robust multi-agent reinforcement learning method based on adversarial domain randomization for real-world dual-uav cooperation. *IEEE Transactions on Intelligent Vehicles*, 2023. 32

291 Ho-Bin Choi, Ju-Bong Kim, Youn-Hee Han, Se-Won Oh, and Kwihoon Kim. Marl-based cooperative multi-agv control in warehouse systems. *IEEE Access*, 10:100478–100488, 2022. 32

292 Yanchang Liang, Xiaowei Zhao, and Li Sun. A multiagent reinforcement learning approach for wind farm frequency control. *IEEE Transactions on Industrial Informatics*, 19(2):1725–1734, 2022. 32

293 Zhenhan Huang and Fumihide Tanaka. Correction: Mspm: A modularized and scalable multi-agent reinforcement learning-based system for financial portfolio management. *PLOS ONE*, 17(3):e0265924, 2022. 32

294 Ali Shavandi and Majid Khedmati. A multi-agent deep reinforcement learning framework for algorithmic trading in financial markets. *Expert Systems with Applications*, 208:118124, 2022. 32

295 Yuling Huang, Chujin Zhou, Kai Cui, and Xiaoping Lu. A multi-agent reinforcement learning framework for optimizing financial trading strategies based on timesnet. *Expert Systems with Applications*, page 121502, 2023. 32

296 Chao Wen, Miao Xu, Zhilin Zhang, Zhenzhe Zheng, Yuhui Wang, Xiangyu Liu, Yu Rong, Dong Xie, Xiaoyang Tan, Chuan Yu, et al. A cooperative-competitive multi-agent framework for auto-bidding in online advertising. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1129–1139, 2022. 32

297 Yuchen Fang, Zhenggang Tang, Kan Ren, Weiqing Liu, Li Zhao, Jiang Bian, Dongsheng Li, Weinan Zhang, Yong Yu, and Tie-Yan Liu. Learning multi-agent intention-aware communication for optimal multi-order execution in finance. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4003–4012, 2023. 32

298 Gian Carlo Cardarilli, Luca Di Nunzio, Rocco Fazzolari, Daniele Giardino, Marco Re, Andrea Ricci, and Sergio Spano. An fpga-based multi-agent reinforcement learning timing synchronizer. *Computers and Electrical Engineering*, 99:107749, 2022. 32

299 Jing-Cheng Shi, Yang Yu, Qing Da, Shi-Yong Chen, and An-Xiang Zeng. Virtual-taobao: Virtualizing real-world online retail environment for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4902–4909, 2019. 32

300 Zool Hilmi Ismail, Nohaidda Sariff, and E Gorrostieta Hurtado. A survey and analysis of cooperative multi-agent robot systems: challenges and directions. *Applications of Mobile Robots*, pages 8–14, 2018. 32

301 Ammar Abdul Ameer Rasheed, Mohammed Najm Abdullah, and Ahmed Sabah Al-Araji. A review of multi-agent mobile robot systems applications. *International Journal of Electrical & Computer Engineering (2088-8708)*, 12(4), 2022. 32

302 Abhinav Dahiya, Alexander M Aroyo, Kerstin Dautenhahn, and Stephen L Smith. A survey of multi-agent human–robot interaction systems. *Robotics and Autonomous Systems*, 161:104335, 2023. 32

303 Paul Maria Scheikl, Balázs Gyenes, Tornike Davitashvili, Rayan Younis, André Schulze, Beat P Müller-Stich, Gerhard Neumann, Martin Wagner, and Franziska Mathis-Ullrich. Cooperative assistance in robotic surgery through multi-agent reinforcement learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1859–1864, 2021. 32

304 Zhaozhi Wang, Kefan Su, Jian Zhang, Huizhu Jia, Qixiang Ye, Xiaodong Xie, and Zongqing Lu. Multi-agent automated machine learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11960–11969, 2023. 32

305 Haiyang Wang, Wenguan Wang, Xizhou Zhu, Jifeng Dai, and Liwei Wang. Collaborative visual navigation. *preprint arXiv:2107.01151*, 2021. 32

306 Zeyu Fang, Jian Zhao, Mingyu Yang, Wengang Zhou, Zhenbo Lu, and Houqiang Li. Coordinate-aligned multi-camera collaboration for active multi-object tracking. *preprint arXiv:2202.10881*, 2022. 32

307 Shiqi Lin, Tao Yu, Ruoyu Feng, Xin Li, Xiaoyuan Yu, Lei Xiao, and Zhibo Chen. Local patch autoaugment with multi-agent collaboration. *IEEE Transactions on Multimedia*, 2023. 32

308 Xinyuan Zhang, Cong Zhao, Feixiong Liao, Xinghua Li, and Yuchuan Du. Online parking assignment in an environment of partially connected vehicles: A multi-agent deep reinforcement learning approach. *Transportation Research Part C: Emerging Technologies*, 138:103624, 2022. 32

309 Yu Sui and Shiming Song. A multi-agent reinforcement learning framework for lithium-ion battery scheduling problems. *Energies*, 13(8):1982, 2020. 32

310 Xiaohan Wang, Lin Zhang, Tingyu Lin, Chun Zhao, Kunyu Wang, and Zhen Chen. Solving job scheduling problems in a resource preemption environment with multi-agent reinforcement learning. *Robotics and Computer-Integrated Manufacturing*, 77:102324, 2022. 32

311 Tong Zhou, Dunbing Tang, Haihua Zhu, and Zequn Zhang. Multi-agent reinforcement learning for online scheduling in smart factories. *Robotics and Computer-Integrated Manufacturing*, 72:102202, 2021. 32

312 Elie Kadoche, Sébastien Gourvénec, Maxime Pallud, and Tanguy Levent. Marlyc: Multi-agent reinforcement learning yaw control. *Renewable Energy*, 217:119129, 2023. 32

313 Yinfeng Yu, Changan Chen, Lele Cao, Fangkai Yang, and Fuchun Sun. Measuring acoustics with collaborative multiple agents. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 335–343, 2023. 32

314 Xu Xu, Youwei Jia, Yan Xu, Zhao Xu, Songjian Chai, and Chun Sing Lai. A multi-agent reinforcement learning-based data-driven method for home energy management. *IEEE Transactions on Smart Grid*, 11(4):3201–3211, 2020. 32

315 Haoyu Zhou, Haifeng Zhang, Yushan Zhou, Xinchao Wang, and Wenxin Li. Botzone: an online multi-agent competitive platform for ai education. In *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, pages 33–38, 2018. 32

316 Yinda Chen, Wei Huang, Shenglong Zhou, Qi Chen, and Zhiwei Xiong. Self-supervised neuron segmentation with multi-agent reinforcement learning. pages 609–617, 2023. 32

317 Hanane Allioui, Mazin Abed Mohammed, Narjes Benameur, Belal Al-Khateeb, Karrar Hameed Abdulkareem, Begonya Garcia-Zapirain, Robertas Damaševičius, and Rytis Maskeliūnas. A multi-agent deep reinforcement learning approach for enhancement of covid-19 ct image segmentation. *Journal of personalized medicine*, 12(2):309, 2022. 32

318 Zihao Gong, Yang Xu, and Delin Luo. Uav cooperative air combat maneuvering confrontation based on multi-agent reinforcement learning. *Unmanned Systems*, 11(03):273–286, 2023. 32

319 Shaowei Li, Yongchao Wang, Yaoming Zhou, Yuhong Jia, Hanyue Shi, Fan Yang, and Chaoyue Zhang. Multi-uav cooperative air combat decision-making based on multi-agent double-soft actor-critic. *Aerospace*, 10(7):574, 2023. 32

320 Lixing Liu, Nikolos Gurney, Kyle McCullough, and Volkan Ustun. Graph neural network based behavior prediction to support multi-agent reinforcement learning in military training simulations. In *2021 Winter Simulation Conference*, pages 1–12. IEEE, 2021. 32

321 Anjon Basak, Erin G Zaroukian, Kevin Corder, Rolando Fernandez, Christopher D Hsu, Piyush K Sharma, Nicholas R Waytowich, and Derrik E Asher. Utility of doctrine with multi-agent rl for military engagements. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV*, volume 12113, pages 609–628. SPIE, 2022. 32

322 KONG Weiren, ZHOU Deyun, Kai Zhang, and YANG Zhen. Air combat autonomous maneuver decision for one-on-one within visual range engagement base on robust multi-agent reinforcement learning. In *2020 IEEE 16th International Conference on Control & Automation (ICCA)*, pages 506–512. IEEE, 2020. 32

323 Haiyin Piao, Yue Han, Shaoming He, Chao Yu, Songyuan Fan, Yaqing Hou, Chengchao Bai, and Li Mo. Spatio-temporal relationship cognitive learning for multi-robot air combat. *IEEE Transactions on Cognitive and Developmental Systems*, 2023. 32

324 Tianrui Jiang, Dongye Zhuang, and Haibin Xie. Anti-drone policy learning based on self-attention multi-agent deterministic policy gradient. In *International Conference on Autonomous Unmanned Systems*, pages 2277–2289, 2021. 32

325 Longfei Yue, Rennong Yang, Jialiang Zuo, Ying Zhang, Qiuni Li, and Yijie Zhang. Unmanned aerial vehicle swarm cooperative decision-making for sead mission: A hierarchical multiagent reinforcement learning approach. *IEEE Access*, 10:92177–92191, 2022. 32

326 Zhang Jiandong, YANG Qiming, SHI Guoqing, LU Yi, and WU Yong. Uav cooperative air combat maneuver decision based on multi-agent reinforcement learning. *Journal of Systems Engineering and Electronics*, 32(6):1421–1438, 2021. 32

327 Wei-ren Kong, De-yun Zhou, Yong-jie Du, Ying Zhou, and Yi-yang Zhao. Hierarchical multi-agent reinforcement learning for multi-aircraft close-range air combat. *IET Control Theory & Applications*, 17(13):1840–1862, 2023. 32

328 Zhixiao Sun, Haiyin Piao, Zhen Yang, Yiyang Zhao, Guang Zhan, Deyun Zhou, Guanglei Meng, Hechang Chen, Xing Chen, Bohao Qu, et al. Multi-agent hierarchical policy gradient for air combat tactics emergence via self-play. *Engineering Applications of Artificial Intelligence*, 98:104112, 2021. 32

329 Zhixiao Sun, Huahua Wu, Yandong Shi, Xiangchao Yu, Yifan Gao, Wenbin Pei, Zhen Yang, Haiyin Piao, and Yaqing Hou. Multi-agent air combat with two-stage graph-attention communication. *Neural Computing and Applications*, 35(27):19765–19781, 2023. 32

330 Gráinne Conole. *Designing for Learning in An Open World*, volume 4. Springer Science & Business Media, 2012. 33

331 Asiiah Song. A little taxonomy of open-endedness. In *ICLR Workshop on Agent Learning in Open-Endedness*, 2022. 33

332 Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020. 33

333 Bo-Jian Hou, Lijun Zhang, and Zhi-Hua Zhou. Learning with feature evolvable streams. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2602–2615, 2019. 33

334 Djordje Grbic, Rasmus Berg Palm, Elias Najarro, Claire Glanois, and Sebastian Risi. Evocraft: A new challenge for open-endedness. In *Applications of Evolutionary Computation*, pages 325–340. Springer, 2021. 33

335 David Balduzzi, Marta Garnelo, Yoram Bachrach, Wojciech Czarnecki, Julien Perolat, Max Jaderberg, and Thore Graepel. Open-ended learning in symmetric zero-sum games. In *Proceedings of the International Conference on Machine Learning*, pages 434–443, 2019. 33

336 Haoqi Yuan, Chi Zhang, Hongcheng Wang, Feiyang Xie, Penglin Cai, Hao Dong, and Zongqing Lu. Plan4mc: Skill reinforcement learning and planning for open-world minecraft tasks. *preprint arXiv:2303.16563*, 2023. 33

337 Robert Meier and Asier Mujika. Open-ended reinforcement learning with neural reward functions. In *Advances in Neural Information Processing Systems*, pages 2465–2479, 2022. 33

338 Michael Matthews, Mikayel Samvelyan, Jack Parker-Holder, Edward Grefenstette, and Tim Rocktäschel. Skillhack: A benchmark for skill transfer in open-ended reinforcement learning. In *ICLR Workshop on Agent Learning in Open-Endedness*, 2022. 33

339 Open Ended Learning Team, Adam Stooke, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, et al. Open-ended learning leads to generally capable agents. *preprint arXiv:2107.12808*, 2021. 33

340 Adaptive Agent Team, Jakob Bauer, Kate Baumli, Satinder Baveja, Feryal Behbahani, Avishkar Bhoopchand, Nathalie Bradley-Schmieg, Michael Chang, Natalie Clay, Adrian Collister, et al. Human-timescale adaptation in an open-ended task space. *preprint arXiv:2301.07608*, 2023. 33

341 Yiqin Yang, Xiaoteng Ma, Chenghao Li, Zewu Zheng, Qiyuan Zhang, Gao Huang, Jun Yang, and Qianchuan Zhao. Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 10299–10312, 2021. 34, 36

342 Jiechuan Jiang and Zongqing Lu. Offline decentralized multi-agent reinforcement learning. *preprint arXiv:2108.01832*, 2021. 34, 36

343 Qi Tian, Kun Kuang, Furui Liu, and Baoxiang Wang. Learning from good trajectories in offline multi-agent reinforcement learning. *preprint arXiv:2211.15612*, 2022. 34, 36

344 Linghui Meng, Muning Wen, Yaodong Yang, Chenyang Le, Xiyun Li, Weinan Zhang, Ying Wen, Haifeng Zhang, Jun Wang, and Bo Xu. Offline pre-trained multi-agent decision transformer: One big sequence model conquers all starcraftii tasks. *preprint arXiv:2112.02845*, 2021. 34, 35, 36, 42

345 Xiangsen Wang and Xianyuan Zhan. Offline multi-agent reinforcement learning with coupled value factorization. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 2781–2783, 2023. 34, 36

346 Jianzhun Shao, Yun Qu, Chen Chen, Hongchang Zhang, and Xiangyang Ji. Counterfactual conservative q learning for offline multi-agent reinforcement learning. *preprint arXiv:2309.12696*, 2023. 34

347 Shayegan Omidshafiei, Dong-Ki Kim, Miao Liu, Gerald Tesauro, Matthew Riemer, Christopher Amato, Murray Campbell, and Jonathan P. How. Learning to teach in cooperative multiagent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6128–6136, 2019. 34, 37

348 Tianpei Yang, Weixun Wang, Hongyao Tang, Jianye Hao, Zhaopeng Meng, Hangyu Mao, Dong Li, Wulong Liu, Yingfeng Chen, Yujing Hu, et al. An efficient transfer learning framework for multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 17037–17048, 2021. 34, 37

349 Qian Long, Zihan Zhou, Abhinav Gupta, Fei Fang, Yi Wu, and Xiaolong Wang. Evolutionary population curriculum for scaling multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2019. 34, 37, 39

350 Anuj Mahajan, Mikayel Samvelyan, Tarun Gupta, Benjamin Ellis, Mingfei Sun, Tim Rocktäschel, and Shimon Whiteson. Generalization in cooperative multi-agent systems. *preprint arXiv:2202.00104*, 2022. 34, 37

351 Rongjun Qin, Feng Chen, Tonghan Wang, Lei Yuan, Xiaoran Wu, Zongzhang Zhang, Chongjie Zhang, and Yang Yu. Multi-agent policy transfer via task relationship modeling. *preprint arXiv:2203.04482*, 2022. 34, 37

352 Hadi Nekoei, Akilesh Badrinaaraayanan, Aaron Courville, and Sarath Chandar. Continuous coordination as a realistic scenario for lifelong learning. In *Proceedings of the International Conference on Machine Learning*, pages 8016–8024, 2021. 34, 38

353 Lei Yuan, Lihe Li, Ziqian Zhang, Fuxiang Zhang, Cong Guan, and Yang Yu. Multi-agent continual coordination via progressive task contextualization. *preprint arXiv:2305.13937*, 2023. 34, 35, 38, 41

354 Lei Yuan, Lihe Li, Ziqian Zhang, Feng Chen, Tianyi Zhang, Cong Guan, Yang Yu, and Zhi-Hua Zhou. Learning to coordinate with anyone. *preprint arXiv:2309.12633*, 2023. 34, 38, 41

355 Somdeb Majumdar, Shauharda Khadka, Santiago Miret, Stephen McAleer, and Kagan Tumer. Evolutionary reinforcement learning for sample-efficient multiagent coordination. In *Proceedings of the International Conference on Machine Learning*, pages 6651–6660, 2020. 34, 38

356 Gaurav Dixit and Kagan Tumer. Balancing teams with quality-diversity for heterogeneous multiagent coordination. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 236–239, 2022. 34, 38

357 Gaurav Dixit, Everardo Gonzalez, and Kagan Tumer. Diversifying behaviors for learning in asymmetric multiagent systems. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 350–358, 2022. 34, 38

358 Lei Yuan, Ziqian Zhang, Ke Xue, Hao Yin, Feng Chen, Cong Guan, Lihe Li, Chao Qian, and Yang Yu. Robust multi-agent coordination via evolutionary generation of auxiliary adversarial attackers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11753–11762, 2023. 34, 39

359 Chenghe Wang, Yuhang Ran, Lei Yuan, Yang Yu, and Zongzhang Zhang. Robust multi-agent reinforcement learning against adversaries on observation. 2022. 34, 39

360 Ziyan Wang, Yali Du, Aivar Sootla, Haitham Bou Ammar, and Jun Wang. CAMA: A new framework for safe multi-agent reinforcement learning using constraint augmentation, 2023. 34, 40

361 Zengyi Qin, Kaiqing Zhang, Yuxiao Chen, Jingkai Chen, and Chuchu Fan. Learning safe multi-agent control with decentralized neural barrier certificates. In *International Conference on Learning Representations*, 2020. 34, 39, 40

362 Ingy ElSayed-Aly, Suda Bharadwaj, Christopher Amato, Rüdiger Ehlers, Ufuk Topcu, and Lu Feng. Safe multi-agent reinforcement learning via shielding. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 483–491, 2021. 34, 40

363 Daniel Melcer, Christopher Amato, and Stavros Tripakis. Shield decentralization for safe multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 13367–13379, 2022. 34, 40

364 Wei-Fang Sun, Cheng-Kuang Lee, and Chun-Yi Lee. Dfac framework: Factorizing the value function via quantile mixture for multi-agent distributional q-learning. In *Proceedings of the International Conference on Machine Learning*, pages 9945–9954, 2021. 34, 40

365 Wei Qiu, Xinrun Wang, Runsheng Yu, Rundong Wang, Xu He, Bo An, Svetlana Obraztsova, and Zinovi Rabinovich. Rmix: Learning risk-sensitive policies for cooperative reinforcement learning agents. In *Advances in Neural Information Processing Systems*, pages 23049–23062, 2021. 34, 40

366 Jifeng Hu, Yanchao Sun, Hechang Chen, Sili Huang, Yi Chang, Lichao Sun, et al. Distributional reward estimation for

effective multi-agent deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 12619–12632, 2022. 34, 40

367 Kyunghwan Son, Junsu Kim, Sungsoo Ahn, Roben D Delos Reyes, Yung Yi, and Jinwoo Shin. Disentangling sources of risk for distributional multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 20347–20368, 2022. 34, 40

368 Ziyi Liu and Yongchun Fang. Learning adaptable risk-sensitive policies to coordinate in multi-agent general-sum games. *arXiv preprint arXiv:2303.07850*, 2023. 34, 40

369 Oliver Slumbers, David Henry Mguni, Stefano B Blumberg, Stephen Marcus Mcaleer, Yaodong Yang, and Jun Wang. A game-theoretic framework for managing risk in multi-agent systems. In *Proceedings of the International Conference on Machine Learning*, pages 32059–32087, 2023. 34, 40

370 Peter Stone, Gal A. Kaminka, Sarit Kraus, and Jeffrey S. Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1504–1509, 2010. 35, 41

371 Muthukumaran Chandrasekaran, A. Eck, Prashant Doshi, and Leen-Kiat Soh. Individual planning in open and typed agent systems. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 82–91, 2016. 35, 41

372 Pengjie Gu, Mengchen Zhao, Jianye Hao, and Bo An. Online ad hoc teamwork under partial observability. In *International Conference on Learning Representations*, 2022. 35, 41

373 Muhammad A Rahman, Niklas Hopner, Filippos Christianos, and Stefano V Albrecht. Towards open ad hoc teamwork using graph-based policy learning. In *Proceedings of the International Conference on Machine Learning*, pages 8776–8786, 2021. 35, 41

374 Arrasy Rahman, Elliot Fosong, Ignacio Carlucho, and Stefano V Albrecht. Generating teammates for training robust ad hoc teamwork agents via best-response diversity. *Transactions on Machine Learning Research*, 2023. 35, 41

375 Arrasy Rahman, Jiaxun Cui, and Peter Stone. Minimum coverage sets for training robust ad hoc teamwork agents. *preprint arXiv:2308.09595*, 2023. 35, 41

376 João G Ribeiro, Gonçalo Rodrigues, Alberto Sardinha, and Francisco S Melo. Teamster: Model-based reinforcement learning for ad hoc teamwork. *Artificial Intelligence*, page 104013, 2023. 35, 41

377 DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. Collaborating with humans without human data. In *Advances in Neural Information Processing Systems*, pages 14502–14515, 2021. 35, 41, 42

378 Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. Trajectory diversity for zero-shot coordination. In *Proceedings of the International Conference on Machine Learning*, pages 7204–7213, 2021. 35, 41

379 Ke Xue, Yutong Wang, Lei Yuan, Cong Guan, Chao Qian, and Yang Yu. Heterogeneous multi-agent zero-shot coordination by coevolution. *preprint arXiv:2208.04957*, 2022. 35, 41

380 Hao Ding, Chengxing Jia, Cong Guan, Feng Chen, Lei Yuan, Zongzhang Zhang, and Yang Yu. Coordination scheme probing for generalizable multi-agent reinforcement learning, 2023. 35, 41

381 Rujikorn Charakorn, Poramate Manoonpong, and Nat Dilokthanakul. Generating diverse cooperative agents by learning incompatible policies. In *International Conference on Learning Representations*, 2023. 35, 41

382 Chao Yu, Jiaxuan Gao, Weilin Liu, Botian Xu, Hao Tang, Jiaqi Yang, Yu Wang, and Yi Wu. Learning zero-shot cooperation with humans, assuming humans are biased. In *International Conference on Learning Representations*, 2023. 35, 41, 42

383 Hadi Nekoei, Xutong Zhao, Janarthanan Rajendran, Miao Liu, and Sarath Chandar. Towards few-shot coordination: Revisiting ad-hoc teamplay challenge in the game of hanabi. *preprint arXiv:2308.10284*, 2023. 35, 41

384 Hengyuan Hu, David J Wu, Adam Lerer, Jakob Foerster, and Noam Brown. Human-ai coordination via human-regularized search and learning. *preprint arXiv:2210.05125*, 2022. 35, 42

385 Joey Hong, Anca Dragan, and Sergey Levine. Learning to influence human behavior with offline reinforcement learning. *preprint arXiv:2303.02265*, 2023. 35, 42

386 Sagar Parekh and Dylan P Losey. Learning latent representations to co-adapt to humans. *Autonomous Robots*, pages 1–26, 2023. 35, 42

387 Xingzhou Lou, Jiaxian Guo, Junge Zhang, Jun Wang, Kaiqi Huang, and Yali Du. Pecan: Leveraging policy ensemble for context-aware zero-shot human-ai coordination. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 679–688, 2023. 35

388 Rundong Wang, Weixuan Wang, Xianhan Zeng, Liang Wang, Zhenjie Lian, Yiming Gao, Feiyu Liu, Siqin Li, Xianliang Wang, QIANG FU, Yang Wei, Lanxiao Huang, Longtao Zheng, Zinovi Rabinovich, and Bo An. Multi-agent multi-game entity transformer, 2023. 35, 42

389 Zhengbang Zhu, Minghuan Liu, Liyuan Mao, Bingyi Kang, Minkai Xu, Yong Yu, Stefano Ermon, and Weinan Zhang.

Madiff: Offline multi-agent learning with diffusion models. *preprint arXiv:2305.17330*, 2023. 35, 42

390 Wenhao Li, Dan Qiao, Baoxiang Wang, Xiangfeng Wang, Bo Jin, and Hongyuan Zha. Semantically aligned task decomposition in multi-agent reinforcement learning. *preprint arXiv:2305.10865*, 2023. 35, 43

391 Julien M Hendrickx and Samuel Martin. Open multi-agent systems: Gossiping with random arrivals and departures. In *Proceedings of the IEEE Annual Conference on Decision and Control*, pages 763–768, 2017. 33

392 Jonathan Cohen, Jilles Steeve Dibangoye, and Abdel-Illah Mouaddib. Open decentralized pomdps. In *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence*, pages 977–984, 2017. 35

393 Jonathan Cohen and Abdel-Illah Mouaddib. Monte-carlo planning for team re-formation under uncertainty: Model and properties. In *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence*, pages 458–465, 2018. 35

394 Jonathan Cohen and Abdel-Illah Mouaddib. Power indices for team reformation planning under uncertainty. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 1886–1888, 2019. 35

395 A. Eck, Maulik Shah, Prashant Doshi, and Leen-Kiat Soh. Scalable decision-theoretic planning in open and typed multiagent systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7127–7134, 2020. 35

396 Anirudh Kakarlapudi, Gayathri Anil, Adam Eck, Prashant Doshi, and Leen-Kiat Soh. Decision-theoretic planning with communication in open multiagent systems. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 938–948, 2022. 35

397 Adam Eck, Leen-Kiat Soh, and Prashant Doshi. Decision making in open agent systems. *AI Magazine*, 2023. 35

398 Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *preprint arXiv:2005.01643*, 2020. 36

399 Rafael Figueiredo Prudencio, Marcos ROA Maximo, and Esther Luna Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 36

400 Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the International Conference on Machine Learning*, pages 2052–2062, 2019. 36

401 Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *preprint arXiv:1911.11361*, 2019. 36

402 Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pages 11761–11771, 2019. 36

403 Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1179–1191, 2020. 36

404 Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Finite-sample analysis for decentralized batch multiagent reinforcement learning with networked agents. *IEEE Transactions on Automatic Control*, 66(12):5925–5940, 2021. 36

405 Ling Pan, Longbo Huang, Tengyu Ma, and Huazhe Xu. Plan better amid conservatism: Offline multi-agent reinforcement learning with actor rectification. In *Proceedings of the International Conference on Machine Learning*, pages 17221–17237, 2022. 36

406 Wei-Cheng Tseng, Tsun-Hsuan Wang, Yen-Chen Lin, and Phillip Isola. Offline multi-agent reinforcement learning with knowledge distillation. In *Advances in Neural Information Processing Systems*, pages 226–237, 2022. 36

407 Linghui Meng, Jingqing Ruan, Xuantang Xiong, Xiyun Li, Xi Zhang, Dengpeng Xing, and Bo Xu. M3: Modularization for multi-task and multi-agent offline pre-training. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 1624–1633, 2023. 36

408 Zhuangdi Zhu, Kaixiang Lin, Anil K Jain, and Jiayu Zhou. Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 36

409 Samir Wadhwania, Dong-Ki Kim, Shayegan Omidshafiei, and Jonathan P. How. Policy distillation and value matching in multiagent reinforcement learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 8193–8200, 2019. 37

410 Weixun Wang, Tianpei Yang, Yong Liu, Jianye Hao, Xiaotian Hao, Yujing Hu, Yingfeng Chen, Changjie Fan, and Yang Gao. From few to more: Large-scale dynamic multiagent curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7293–7300, 2020. 37

411 Siyi Hu, Fengda Zhu, Xiaojun Chang, and Xiaodan Liang. UPDeT: Universal multi-agent reinforcement learning via policy decoupling with transformers. In *International Conference on Learning Representations*, 2021. 37

412 Tianze Zhou, Fubiao Zhang, Kun Shao, Kai Li, Wenhan Huang, Jun Luo, Weixun Wang, Yaodong Yang, Hangyu Mao, Bin

Wang, et al. Cooperative multi-agent transfer learning with level-adaptive credit assignment. *preprint arXiv:2106.00517*, 2021. 37

413 Haobin Shi, Jingchen Li, Jiahui Mao, and Kao-Shing Hwang. Lateral transfer learning for multiagent reinforcement learning. *IEEE transactions on cybernetics*, 53(3):1699–1711, 2023. 37

414 Dhireesha Kudithipudi, Mario Aguilar-Simon, Jonathan Babb, Maxim Bazhenov, Douglas Blackiston, Josh Bongard, Andrew P Brna, Suraj Chakravarthi Raja, Nick Cheney, Jeff Clune, et al. Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, 4(3):196–210, 2022. 37

415 David Abel, André Barreto, Benjamin Van Roy, Doina Precup, Hado van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. *preprint arXiv:2307.11046*, 2023. 37

416 German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. 37

417 James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 37

418 David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, pages 348–358, 2018. 37

419 Yizhou Huang, Kevin Xie, Homanga Bharadhwaj, and Florian Shkurti. Continual model-based reinforcement learning with hypernetworks. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 799–805, 2021. 38

420 Samuel Kessler, Piotr Miłoś, Jack Parker-Holder, and Stephen J Roberts. The surprising effectiveness of latent world models for continual reinforcement learning. *preprint arXiv:2211.15944*, 2022. 38

421 Zhi Wang, Chunlin Chen, and Daoyi Dong. Lifelong incremental reinforcement learning with online bayesian inference. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8):4003–4016, 2022. 38

422 Samuel Kessler, Jack Parker-Holder, Philip J. Ball, Stefan Zohren, and Stephen J. Roberts. Same state, different task: Continual reinforcement learning without interference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7143–7151, 2022. 38

423 Jean-Baptiste Gaya, Thang Doan, Lucas Caccia, Laure Soulier, Ludovic Denoyer, and Roberta Raileanu. Building a subspace of policies for scalable continual learning. In *International Conference on Learning Representations*, 2023. 38

424 Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018. 38

425 Zhi-Hua Zhou, Yang Yu, and Chao Qian. *Evolutionary learning: Advances in theories and algorithms*. Springer, 2019. 38

426 Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53:659–697, 2015. 38

427 Shayegan Omidshafiei, Christos Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M Czarnecki, Marc Lanctot, Julien Perolat, and Remi Munos. α-rank: Multi-agent evaluation by evolution. *Scientific reports*, 9(1):9937, 2019. 38

428 Takanori Shibata and Toshio Fukuda. Coordination in evolutionary multi-agent-robotic system using fuzzy and genetic algorithm. *Control Engineering Practice*, 2(1):103–111, 1994. 38

429 Pengyi Li, Jianye Hao, Hongyao Tang, Yan Zheng, and Xian Fu. Race: Improve multi-agent reinforcement learning with representation asymmetry and collaborative evolution. In *Proceedings of the International Conference on Machine Learning*, pages 19490–19503, 2023. 38

430 Gaurav Dixit and Kagan Tumer. Learning synergies for multi-objective optimization in asymmetric multiagent systems. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 447–455, 2023. 38

431 Janosch Moos, Kay Hansel, Hany Abdulsamad, Svenja Stark, Debora Clever, and Jan Peters. Robust reinforcement learning: A review of foundations and recent advances. *Machine Learning and Knowledge Extraction*, 4(1):276–315, 2022. 39

432 Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 2817–2826, 2017. 39

433 Eugene Vinitsky, Yuqing Du, Kanaad Parvate, Kathy Jang, Pieter Abbeel, and Alexandre Bayen. Robust reinforcement learning using adversarial populations. *preprint arXiv:2008.01825*, 2020. 39

434 Huan Zhang, Hongge Chen, Duane S Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. In *International Conference on Learning Representations*, 2020. 39

435 Yeeho Song and Jeff Schneider. Robust reinforcement learning via genetic curriculum. In *Proceedings of the International*

Conference on Robotics and Automation, pages 5560–5566, 2022. 39

436 Tuomas Oikarinen, Wang Zhang, Alexandre Megretski, Luca Daniel, and Tsui-Wei Weng. Robust deep reinforcement learning through adversarial loss. In *Advances in Neural Information Processing Systems*, pages 26156–26167, 2021. 39

437 Yanchao Sun, Ruijie Zheng, Yongyuan Liang, and Furong Huang. Who is the strongest enemy? towards optimal and efficient evasion attacks in deep rl. In *International Conference on Learning Representations*, 2021. 39

438 Yongyuan Liang, Yanchao Sun, Ruijie Zheng, and Furong Huang. Efficient adversarial training without attacking: Worst-case-aware robust reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 22547–22561, 2022. 39

439 Michael Everett, Björn Lütjens, and Jonathan P. How. Certifiable robustness to adversarial state uncertainty in deep reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4184–4198, 2022. 39

440 Fan Wu, Linyi Li, Zijian Huang, Yevgeniy Vorobeychik, Ding Zhao, and Bo Li. Crop: Certifying robust policies for reinforcement learning through functional smoothing. In *International Conference on Learning Representations*, 2022. 39

441 Fan Wu, Linyi Li, Huan Zhang, Bhavya Kailkhura, Krishnaram Kenthapadi, Ding Zhao, and Bo Li. Copa: Certifying robust policies for offline reinforcement learning against poisoning attacks. In *International Conference on Learning Representations*, 2021. 39

442 Jieyu Lin, Kristina Dzeparoska, Sai Qian Zhang, Alberto Leon-Garcia, and Nicolas Papernot. On the robustness of cooperative multi-agent reinforcement learning. In *2020 IEEE Security and Privacy Workshops*, pages 62–68, 2020. 39

443 Shihui Li, Yi Wu, Xinyue Cui, Honghua Dong, Fei Fang, and Stuart Russell. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4213–4220, 2019. 39

444 Rupert Mitchell, Jan Blumenkamp, and Amanda Prorok. Gaussian process based message filtering for robust multi-agent cooperation in the presence of adversarial communication. *preprint arXiv:2012.00508*, 2020. 39

445 James Tu, Tsunhsuan Wang, Jingkang Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Adversarial attacks on multi-agent communication. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7768–7777, 2021. 39

446 Thomy Phan, Thomas Gabor, Andreas Sedlmeier, Fabian Ritz, Bernhard Kempter, Cornel Klein, Horst Sauer, Reiner Schmid, Jan Wieghardt, Marc Zeller, et al. Learning and testing resilience in cooperative multi-agent systems. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 1055–1063, 2020. 39

447 Thomy Phan, Lenz Belzner, Thomas Gabor, Andreas Sedlmeier, Fabian Ritz, and Claudia Linnhoff-Popien. Resilient multi-agent reinforcement learning with adversarial value decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11308–11316, 2021. 39

448 Sihong He, Songyang Han, Sanbao Su, Shuo Han, Shaofeng Zou, and Fei Miao. Robust multi-agent reinforcement learning with state uncertainty. *Transactions on Machine Learning Research*, 2023. 39

449 Kalyanmoy Deb, Karthik Sindhya, and Jussi Hakanen. Multi-objective optimization. In *Decision sciences*, pages 161–200. CRC Press, 2016. 39

450 Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48(1):67–113, 2013. 40

451 Chunming Liu, Xin Xu, and Dewen Hu. Multiobjective reinforcement learning: A comprehensive overview. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(3):385–398, 2014. 40

452 Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In *Advances in Neural Information Processing Systems*, pages 14610–14621, 2019. 40

453 Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022. 40

454 Roxana Rădulescu, Patrick Mannion, Diederik M Roijers, and Ann Nowé. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems*, 34(1):10, 2020. 40

455 Ishan Durugkar, Elad Liebman, and Peter Stone. Balancing individual preferences and shared objectives in multiagent reinforcement learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2505–2511, 2020. 40

456 Willem Röpke. Reinforcement learning in multi-objective multi-agent systems. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 2999–3001, 2023. 40

457 Roxana Radulescu, Timothy Verstraeten, Yijie Zhang, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. Opponent

learning awareness and modelling in multi-objective normal form games. *Neural Computing and Applications*, 34(3):1759–1781, 2022. 40

458 Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015. 40

459 Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, and Alois Knoll. A review of safe reinforcement learning: Methods, theory and applications. *preprint arXiv:2205.10330*, 2022. 40

460 Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient generalized lagrangian policy optimization for safe multi-agent reinforcement learning. In *Learning for Dynamics and Control Conference*, pages 315–332, 2023. 40

461 Donghao Ying, Yunkai Zhang, Yuhao Ding, Alec Koppel, and Javad Lavaei. Scalable primal-dual actor-critic method for safe multi-agent rl with general utilities. *preprint arXiv:2305.17568*, 2023. 40

462 Zhili Zhang, Yanchao Sun, Furong Huang, and Fei Miao. Safe and robust multi-agent reinforcement learning for connected autonomous vehicles under state perturbations. *preprint arXiv:2309.11057*, 2023. 40

463 Charles Dawson, Sicun Gao, and Chuchu Fan. Safe control with learned certificates: A survey of neural lyapunov, barrier, and contraction methods for robotics and control. *IEEE Transactions on Robotics*, 2023. 40

464 Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional reinforcement learning*. MIT Press, 2023. 40

465 Noa Agmon and Peter Stone. Leading ad hoc agents in joint action settings with multiple teammates. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 341–348, 2012. 41

466 Stefano V. Albrecht and Peter Stone. Reasoning about hypothetical agent behaviours and their parameters. In *Proceedings of the Conference on Autonomous Agents and MultiAgent Systems*, pages 547–555, 2017. 41

467 Samuel Barrett, Avi Rosenfeld, Sarit Kraus, and Peter Stone. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence*, 242:132–171, 2017. 41

468 Manish Ravula, Shani Alkoby, and Peter Stone. Ad hoc teamwork with behavior switching agents. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 550–556, 2019. 41

469 William Macke, Reuth Mirsky, and Peter Stone. Expected value of communication for planning in ad hoc teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11290–11298, 2021. 41

470 Samuel Barrett, Noa Agmon, Noam Hazon, Sarit Kraus, and Peter Stone. Communicating with unknown teammates. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 1433–1434, 2014. 41

471 Arrasy Rahman, Ignacio Carlucho, Niklas Höpner, and Stefano V Albrecht. A general learning framework for open ad hoc teamwork using graph-based policy learning. *preprint arXiv:2210.05448*, 2022. 41

472 Ted Fujimoto, Samrat Chatterjee, and Auroop Ganguly. Ad hoc teamwork in the presence of adversaries. *preprint arXiv:2208.05071*, 2022. 41

473 Gerald Tesauro. Td-gammon, a self-teaching backgammon program, achieves master-level play. *Neural computation*, 6(2):215–219, 1994. 41

474 David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. 41

475 Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "other-play" for zero-shot coordination. In *Proceedings of the International Conference on Machine Learning*, pages 4399–4410, 2020. 41, 42

476 Johannes Heinrich, Marc Lanctot, and David Silver. Fictitious self-play in extensive-form games. In *Proceedings of the International Conference on Machine Learnin*, pages 805–813, 2015. 41

477 Xihuai Wang, Shao Zhang, Wenhao Zhang, Wentao Dong, Jingxiao Chen, Ying Wen, and Weinan Zhang. Quantifying zero-shot coordination capability with behavior preferring partners. *preprint arXiv:2310.05208*, 2023. 41

478 Rui Zhao, Jinming Song, Hu Haifeng, Yang Gao, Yi Wu, Zhongqian Sun, and Yang Wei. Maximum entropy population based training for zero-shot human-AI coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6145–6153, 2023. 41

479 Darius Muglich, Christian Schroeder de Witt, Elise van der Pol, Shimon Whiteson, and Jakob Foerster. Equivariant networks for zero-shot coordination. In *Advances in Neural Information Processing Systems*, pages 6410–6423, 2022. 41

480 Lebin Yu, Yunbo Qiu, Quanming Yao, Xudong Zhang, and Jian Wang. Improving zero-shot coordination performance based on policy similarity. *preprint arXiv:2302.05063*, 2023. 41

481 Yang Li, Shao Zhang, Jichen Sun, Yali Du, Ying Wen, Xinbing Wang, and Wei Pan. Cooperative open-ended learning framework for zero-shot coordination. *preprint arXiv:2302.04831*, 2023. 41

482 Saurabh Kumar, Aviral Kumar, Sergey Levine, and Chelsea Finn. One solution is not all you need: Few-shot extrapolation via structured maxent RL. In *Advances in Neural Information Processing Systems*, pages 8198–8210, 2020. 41

483 Takayuki Osa, Voot Tangkaratt, and Masashi Sugiyama. Discovering diverse solutions in deep reinforcement learning by maximizing state–action-based mutual information. *Neural Networks*, 152:90–104, 2022. 41

484 Jean-Baptiste Gaya, Laure Soulier, and Ludovic Denoyer. Learning a subspace of policies for online adaptation in reinforcement learning. In *International Conference on Learning Representations*, 2022. 41

485 Arash Ajoudani, Andrea Maria Zanchettin, Serena Ivaldi, Alin Albu-Schäffer, Kazuhiro Kosuge, and Oussama Khatib. Progress and prospects of the human–robot collaboration. *Autonomous Robots*, 42:957–975, 2018. 42

486 Federico Vicentini. Collaborative robotics: a survey. *Journal of Mechanical Design*, 143(4):040802, 2021. 42

487 Niels Van Berkel, Mikael B Skov, and Jesper Kjeldskov. Human-ai interaction: intermittent, continuous, and proactive. *Interactions*, 28(6):67–71, 2021. 42

488 Linda Onnasch and Eileen Roesler. A taxonomy to structure and analyze human–robot interaction. *International Journal of Social Robotics*, 13(4):833–849, 2021. 42

489 Rose E Wang, Sarah A Wu, James A Evans, Joshua B Tenenbaum, David C Parkes, and Max Kleiman-Weiner. Too many cooks: Coordinating multi-agent collaboration through inverse planning. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 2032–2034, 2020. 42

490 Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration. In *International Conference on Learning Representations*, 2020. 42

491 Adam Lerer and Alexander Peysakhovich. Learning existing social conventions via observationally augmented self-play. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 107–114, 2019. 42

492 Mycal Tucker, Yilun Zhou, and Julie Shah. Adversarially guided self-play for adopting social conventions. *preprint arXiv:2001.05994*, 2020. 42

493 Andy Shih, Arjun Sawhney, Jovana Kondic, Stefano Ermon, and Dorsa Sadigh. On the critical role of conventions in adaptive human-ai collaboration. In *International Conference on Learning Representations*, 2020. 42

494 Mengxi Li, Minae Kwon, and Dorsa Sadigh. Influencing leading and following in human–robot teams. *Autonomous Robots*, 45:959–978, 2021. 42

495 Yang Li, Shao Zhang, Jichen Sun, Wenhao Zhang, Yali Du, Ying Wen, Xinbing Wang, and Wei Pan. Tackling cooperative incompatibility for zero-shot human-ai coordination. *preprint arXiv:2306.03034*, 2023. 42

496 Arun Kumar AV, Santu Rana, Alistair Shilton, and Svetha Venkatesh. Human-ai collaborative bayesian optimisation. In *Advances in Neural Information Processing Systems*, pages 16233–16245, 2022. 42

497 Jakob Thumm, Felix Trost, and Matthias Althoff. Human-robot gym: Benchmarking reinforcement learning in human-robot collaboration. *preprint arXiv:2310.06208*, 2023. 42

498 Maxence Hussonnois, Thommen George Karimpanal, and Santu Rana. Controlled diversity with preference: Towards learning a diverse set of desired skills. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 1135–1143, 2023. 42

499 Luyao Yuan, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. In situ bidirectional human-robot value alignment. *Science robotics*, 7(68):eabm4183, 2022. 42

500 Meng Guo and Mathias Bürger. Interactive human-in-the-loop coordination of manipulation skills learned from demonstration. In *2022 International Conference on Robotics and Automation*, pages 7292–7298, 2022. 42

501 Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *preprint arXiv:2303.18223*, 2023. 42, 43

502 Weirui Ye, Yunsheng Zhang, Mengchen Wang, Shengjie Wang, Xianfan Gu, Pieter Abbeel, and Yang Gao. Foundation reinforcement learning: towards embodied generalist agents with foundation prior assistance. *preprint arXiv:2310.02635*, 2023. 42

503 Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. Foundation models for decision making: Problems, methods, and opportunities. *preprint arXiv:2303.04129*, 2023. 42, 43

504 Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *Transactions on Machine Learning Research*, 2022. 42

505 Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *preprint arXiv:2301.04104*, 2023. 42

506 Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in neural information processing systems*, pages 15084–15097, 2021. 42

507 Shengchao Hu, Li Shen, Ya Zhang, Yixin Chen, and Dacheng Tao. On transforming reinforcement learning by transformer: The development trajectory. *preprint arXiv:2212.14164*, 2022. 42

508 Zhuoran Li, Ling Pan, and Longbo Huang. Beyond conservatism: Diffusion policies in offline multi-agent reinforcement learning. *preprint arXiv:2307.01472*, 2023. 42

509 Tao Li, Juan Guevara, Xinghong Xie, and Quanyan Zhu. Self-confirming transformer for locally consistent online adaptation in multi-agent reinforcement learning. *preprint arXiv:2310.04579*, 2023. 42

510 Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *preprint arXiv:2304.03442*, 2023. 42

511 Ziluo Ding, Wanpeng Zhang, Junpeng Yue, Xiangjun Wang, Tiejun Huang, and Zongqing Lu. Entity divider with language grounding in multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 8103–8119, 2023. 43

512 Hengyuan Hu and Dorsa Sadigh. Language instructed reinforcement learning for human-ai coordination. In *Proceedings of the International Conference on Machine Learning*, pages 13584–13598, 2023. 43

513 Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, et al. Proagent: Building proactive cooperative ai with large language models. *preprint arXiv:2308.11339*, 2023. 43

514 Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *preprint arXiv:2305.19118*, 2023. 43

515 Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *preprint arXiv:2308.07201*, 2023. 43

516 Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *preprint arXiv:2308.08155*, 2023. 43

517 Nuoya Xiong, Zhihan Liu, Zhaoran Wang, and Zhuoran Yang. Sample-efficient multi-agent rl: An optimization perspective. *preprint arXiv:2310.06243*, 2023. 43

518 Wanpeng Zhang and Zongqing Lu. Rladapter: Bridging large language models to reinforcement learning in open worlds. *preprint arXiv:2309.17176*, 2023. 43

519 Yuxi Li. Reinforcement learning applications. *preprint arXiv:1908.06973*, 2019. 44

A Survey of Progress on Cooperative Multi-agent Reinforcement Learning in Open Environment

Lei Yuan, Ziqian Zhang, Lihe Li, Cong Guan & Yang Yu*

State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023*, China*
* Corresponding author. E-mail: yuy@nju.edu.cn

Abstract Multi-agent Reinforcement Learning (MARL) has gained widespread attention and made significant progress in various fields in recent years. Among them, cooperative MARL focuses on training teams of agents to collaboratively achieve tasks that a single agent cannot accomplish alone. It has shown tremendous application potential in scenarios such as pathfinding, autonomous driving, active voltage control, and dynamic algorithm configuration. Enhancing coordination efficiency is one of the key research areas in cooperative MARL. Previous methods have mainly focused on operating in simple, static, and closed environments. Driven by the implementation of artificial intelligence technology, there are now some studies in the field of multi-agent cooperation that are beginning to investigate coordination in open environments. These works explore and study situations where the coordination environment for agents may change in various ways. While progress has been made in multiple aspects, mainstream research needs a comprehensive review and summary of this direction. This paper starts with the definition of reinforcement learning and introduces MARL, cooperative MARL, typical methods, and testing environments. It then reviews multi-agent cooperation from closed environments to open settings, classifies relevant works, and describes representative studies within this domain. Finally, it summarizes the strengths and weaknesses of current research and provides prospects for future development and research directions in cooperative MARL in open environments. This aims to attract more researchers to engage in researches and discussions in this emerging field.

Keywords Reinforcement Learning, Multi-agent System, Multi-agent Coordination, Open Environment Machine Learning, Open Environment Multi-agent Coordination

Lei Yuan received the B.Sc. degree in Department of Electronic Engineering in June 2016 from Tsinghua University, and his M.Sc. degree from Chinese Aeronautical Establishment in June 2019. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Nanjing University, Nanjing, China. His current research interests mainly include machine learning, reinforcement learning, and multi-agent reinforcement learning.

Ziqian Zhang received his B.Sc. degree in Brain Science and Artificial Intelligence from Kuang Yaming Honors School, Nanjing University in 2023. He is currently pursuing the M.Sc. degree with the School of Artificial Intelligence, Nanjing University, Nanjing, China. His research interests include multi-agent reinforcement learning and multi-agent system.

Lihe Li received his B.Sc. degree of Engineering from School of Artificial Intelligence, Nanjing University in 2023. He is currently pursuing the M.Sc. degree with the School of Artificial Intelligence, Nanjing University, Nanjing, China. His research interests include multi-agent reinforcement learning and multi-agent system.

Yang Yu received the Ph.D. degree in Computer Science from Nanjing University in 2011, and then joined the LAMDA Group in the Department of Computer Science and Technology of Nanjing University as an Assistant Researcher from 2011, and as an Associate Professor from 2014. He joined the School of Artificial Intelligence of Nanjing University as a Professor from 2019. Prof. Yu's research interest is in machine learning, a subfield of artificial intelligence. Currently, Prof. Yu is working on reinforcement learning in various aspects, including optimization, representation, transfer, etc. Prof. Yu was a recipient of the National Outstanding Doctoral Dissertation Award, the China Computer Federation Outstanding Doctoral Dissertation Award, the PAKDD'08 Best Paper Award, the GECCO'11 Best Paper (Theory Track) and the Microsoft Research Asia Collaborative Research Award. He is an associate editor for Frontiers of Computer Science and an area chair of ACML'17, IJCAI'18, and ICPR'18.