工程科学学报
Chinese Journal of Engineering

# 基于学习机制的多智能体强化学习综述

王若男 董琦

# Multiagent game decision-making method based on the learning mechanism

WANG Ruonan, DONG Qi

引用本文:
王若男, 董琦. 基于学习机制的多智能体强化学习综述[J]. 北科大：工程科学学报, 2024, 46(7): 1251−1268. doi: 10.13374/j.issn2095−9389.2023.08.08.003
WANG Ruonan, DONG Qi. Multiagent game decision−making method based on the learning mechanism[J]. *Chinese Journal of Engineering*, 2024, 46(7): 1251−1268. doi: 10.13374/j.issn2095−9389.2023.08.08.003

在线阅读 View online: https://doi.org/10.13374/j.issn2095−9389.2023.08.08.003

工程科学学报, 第 46 卷, 第 7 期: 1251−1268, 2024 年 7 月
Chinese Journal of Engineering, Vol. 46, No. 7: 1251−1268, July 2024
https://doi.org/10.13374/j.issn2095-9389.2023.08.08.003; http://cje.ustb.edu.cn

# 基于学习机制的多智能体强化学习综述

王若男, 董　琦✉

中国电子科学研究院, 北京 100041
✉通信作者, E-mail: dongqiouc@126.com

**摘　要**　强化学习作为人工智能领域的重要分支, 以其在多智能体系统决策中的卓越表现, 成为当前主流方法. 然而, 传统的多智能体强化学习算法在面对维度爆炸、训练样本稀缺和难以迁移等方面仍然存在困难. 为了克服这些挑战并提升算法性能, 本文从学习机制的角度入手, 深入研究学习机制与强化学习的深度融合, 以推动多智能体强化学习算法的发展. 首先, 介绍了多智能体强化学习算法的基本原理、发展历程以及算法所面临的难点. 随后, 引入了基于学习机制的多智能体强化学习方法这一种新兴方向. 这些学习机制, 如元学习和迁移学习, 被证明可以有效提升多智能体的学习速度, 并缓解维度爆炸等问题. 按照课程学习、演化博弈、元学习、分层学习、迁移学习等学习机制在多智能体强化学习中的应用进行了综述, 通过罗列这些方法的研究成果, 论述了各种方法的局限性, 并提出了未来改进的方向. 总结了这类融合算法在实际应用中取得的提升成果和实际应用, 具体列举了基于学习机制的多智能体强化学习算法在交通控制、游戏领域的实际应用案例. 同时, 对这类融合算法未来在理论、算法和应用方面的发展方向进行了深入分析. 这涵盖了对新颖理论的探索、算法性能的进一步优化, 以及在更广泛领域中的推广应用. 通过这样的综述和分析, 为未来多智能体强化学习算法的研究方向和实际应用提供了有益的参考.

**关键词**　强化学习; 多智能体博弈; 学习机制; 课程学习; 演化强化学习
**分类号**　TP312

# Multiagent game decision-making method based on the learning mechanism

*WANG Ruonan*, *DONG Qi*✉

China Academy of Electronics Science, Beijing 100041, China
✉Corresponding author, E-mail: dongqiouc@126.com

**ABSTRACT**　Reinforcement learning, a cornerstone in the expansive landscape of artificial intelligence, has asserted its dominance as the prevailing methodology in contemporary multiagent system decision-making because of its formidable efficacy. However, the path to the zenith of algorithmic excellence is fraught with challenges intrinsic to traditional multiagent reinforcement learning algorithms, such as dimensionality explosion, scarcity of training samples, and the labyrinthine nature of migration processes. In a concerted effort to surmount these formidable challenges and propel the evolution of algorithmic prowess, this paper unfurls its inquiry from the perspective of learning mechanisms and undertakes an exhaustive exploration of the symbiotic integration of learning mechanisms and reinforcement learning. At the inception of this scholarly expedition, we meticulously delineate the rudimentary principles underpinning multiagent algorithms, present a historical trajectory tracing their developmental evolution, and cast a discerning eye upon the salient challenges that have been formidable impediments in their trajectory. The ensuing narrative charts a course into the avant-garde realm of multiagent reinforcement learning methods anchored in learning mechanisms, a paradigmatic shift that emerges as an innovative frontier in the field. Among these learning mechanisms, meta-learning and transfer learning are empirically validated as useful instruments in hastening the

**收稿日期:** 2023−08−08
**基金项目:** 网络空间安全态势感知与评估安徽省重点实验室开放课题资助项目(CSSAE-2021-003); 国家自然科学基金项目青年科学基金资助项目(62206018)

learning trajectory of multiagent systems and simultaneously mitigating the intricate challenges posed by dimensionality explosion. This paper assumes the role of a sagacious guide through the labyrinthine landscape of multiagent reinforcement learning, focusing on the manifold applications of learning mechanisms across diverse domains. A comprehensive review delineates the impact of learning mechanisms in curriculum learning, evolutionary games, meta-learning, hierarchical learning, and transfer learning. The research outcomes within these thematic realms are methodically cataloged, with a discerning eye cast upon the limitations inherent in each methodology and erudite propositions for the trajectory of future improvements. The discourse pivots toward synthesizing advancements and accomplishments wrought by fusion algorithms in practical milieus. This paper meticulously examines the transformative impact of fusion algorithms in real-world applications, with a detailed exposition of their deployment in domains as diverse as traffic control and gaming. Simultaneously, an incisive analysis charting the future trajectory of fusion algorithms is conducted. This prediction encompasses exploring nascent theories, refining algorithmic efficacy, and expanding dissemination and application across a broader spectrum of domains. Through this scholarly odyssey, this paper provides an invaluable compass for navigating the uncharted waters of future research endeavors and the judicious deployment of multiagent reinforcement learning algorithms in pragmatic scenarios.

**KEY WORDS** reinforcement learning; multiagent game; learning mechanism; curriculum learning; evolutionary reinforcement learning

近年来, 随着计算能力和新一代信息技术的飞速发展, 人工智能正处于由感知向认知决策跨越的阶段, 越来越多的人工智能研究者开始关注“决策智能”. 目前, 自主智能的研究对象已经从个体扩展到群体, 群体智能决策成为一个研究重点. 20 世纪 70 年代以来, 研究者对多智能体决策领域展开了众多的研究, 目的在于建立一个拥有自主学习能力的多智能体决策系统[1−2].

多智能体强化学习算法 (Multi-agent reinforcement learning, MARL) 是多智能体博弈决策领域的主流方法. MARL 中的智能体通过不断与环境和其他智能体进行交互获得信息, 学习如何在自身利益和群体收益之间进行权衡, 以达到最优策略. 国内外已有的综述从不同的角度讨论了 MARL 算法. 例如, Gronauer 等[3] 和 Wang 等[4] 清晰阐述了当前 MARL 的研究现状和面对的挑战; Yang 和 Wang[5] 从博弈论的角度梳理了近年来出现的强化学习算法,总结当前博弈强化学习算法的重难点; Zhang 等[6] 重点介绍了 MARL 理论的几个新角度和分类, 包括广泛形式博弈中的学习、具有网络智能体的去中心化 MARL、均场体系中的 MARL 等; Oroojlooy 等[7] 重点介绍了建模和解决合作多智能体强化学习问题的五种常见方法以及相关的应用; Hernandez-Leal 等[8] 从涌现行为、沟通和合作学习的角度综述了多智能体决策方法; Da Silva[9] 回顾了多智能体强化学习中的知识重用自治方法; Bloembergen 等[10] 分析了进化动力学在智能决策中的启发等.

随着计算能力的发展, MARL 在一系列的领域中逐渐实现了人类甚至超人类级别的性能. 但是 MARL 仍然存在的一些尚未解决的难题, 首先状态空间和动作空间可能会随着智能体数量的增加呈指数级增长, 难以在可接受的时间范围内学习到最优策略; 其次深度强化学习训练需要大量的训练数据, 通常难以在较少经验的情况下在广泛的任务中获得合理性; 除此以外, 强化学习系统通常专注于一个受限的任务领域, 无法灵活地适应变化的任务条件等等[11−13].

因此, 本文从学习机制的角度出发, 集中调研了课程学习、演化学习、元学习、分层学习和迁移学习五类学习机制与多智能体强化学习的结合. 这些机制通过引入额外的学习过程和或计算方法, 从不同方面提升了多智能体算法性能, 如有效解决维度爆炸带来的计算问题、提升算法在少样本任务上的表现、引导多智能快速适应于新的场景和任务等等. 本文的主要贡献如下:

(1) 创造性地从学习机制的角度出发, 介绍了一系列基于学习机制的多智能体强化学习算法的机理. 学习机制的融入提升了多智能体强化学习算法的性能, 为应对多智能体强化学习算法的固有挑战提供了新的方案.

(2) 分析基于学习机制的多智能体强化学习算法的原理和适用场景, 并罗列了每类融合算法的研究成果和发展历程, 对算法的优势和局限性进行总结, 并给出未来可改进的方向, 对不同类算法的调查结果以表格的形式呈现.

(3) 介绍了基于学习机制的多智能体强化学习算法实际应用, 并给出了这一融合算法在理论、方法、场景的未来发展方向, 以期在未来解决更加复杂和更加具有挑战性的多智能体博弈决策问题.

## 1 背景知识

### 1.1 单智能体强化学习

马尔可夫决策过程(Markov decision process, MDP)是单智能体决策的求解模型. 智能体与环境之间不断交互, 获得随机性的策略和回报. MDP 的要素为五元组 $<S,A,R,P,\rho_0>$, $S$ 是状态空间; $A$ 是动作空间; 奖励值 $R:S\times A\times S\to \mathbf{R}$; $P$: $S\times A\to \mathrm{P}(S)$ 是状态转移的规则; $\rho_0$ 是开始状态的分布. MDP 状态的转移关系由当前状态 $s_t$ 与输入的行动 $a_t$ 共同决定. MDP 的求解目标是找到预期收益最大对应的策略, 一般用动作值函数 $Q(s,a)$ 来形式化地表征收益的期望, $Q(s,a)$ 是在某个状态 $s$ 下采取动作 $a$ 后, 再根据策略 $\pi$ 进行一系列动作后得到的累计奖励的期望为:

$$Q_\pi(s,a)=E_{\tau\sim\pi}[R(\tau)|s_0=s,a_0=a] \tag{1}$$

最优动作值函数表示为 $Q^*(s,a)$, 从而可以求解出最优动作, 一系列最优动作构建最优策略. Q-Learning 是一种经典的强化学习方式, 它以表格的形式存储 $Q(s,a)$, 算法迭代更新 $Q$ 表格, 最终得到 $Q^*(s,a)$, 从而求解最优策略. 基于值的强化学习算法如深度 $Q$ 网络(Deep Q net, DQN)利用深度卷积网络来逼近 $Q(s,a)$, 并将交互数据以记忆单元 $<s,a,r,s'>$ 的形式存放, 然后以随机小样本采样进行网络训练和参数更新. 基于策略(Policy-based)的算法跨越价值函数, 直接搜索最佳策略, 通过最大化累计回报来更新策略参数. 这种方法下的策略显示地表示为,对预计收益进行梯度下降优化. 执行者-评价器(Actor-critic, AC)在策略梯度中引入动作值函数对选取的动作进行评分, 然后执行者根据评分调整选动作的概率.

### 1.2 多智能体强化学习

随机博弈 (Stochastic games, SG) 是 MDP 的自然衍生, 是多智能体强化学习算法的求解模型,可由元组表示: $(n,S,A_1,A_2,\cdots,A_n,T,r_1,r_2,\cdots,r_n,\gamma)$. 其中 $n$ 为智能体的数目, $S$ 是联合状态的空间, $A_i$ 为第 $i$ 个智能体的动作空间, 联合动作空间可以被定义为 $A=A_1\times A_2\times\cdots\times A_n$; $T$ 是联合状态转移概率, $T:S\times A\times S\to[0,1]$, 它决定了在执行联合动作 $a\in A$ 的情况下, 由状态 $s\in S$ 转移到下一个状态 $s\in S'$ 的概率分布; $r_i:S\times A\times S\to\mathbf{R}^n$ 为第 $i$ 个智能体的回报. 在随机博弈中, 智能体的下一状态的奖励和状态跟智能体的联合动作有关[13]. MARL 中, 每个智能体都与环境交互获取奖励值, 从而获得该环境下的最优策略. MARL 领域存在着一些固有挑战, 智能体数量增加带来的维度爆炸的问题, 最终导致算法收敛困难、计算复杂; 多智能体中奖励机制设置复杂, 往往存在着稀疏奖励的情况; 学习到的模型泛化性较差, 难以迁移到不同的场景中等等. 如经典算法多智能体深度确定性策略 (Multi-agent deep deterministic policy gradient, MADDPG) 在处理非平稳的环境中会很难收敛, 同时 MADDPG 将其他智能体的信息作为输入状态, 导致了算法无法应用于大规模多智能体场景中[14]. 反事实多智能体策略梯度 (Counterfactual multi-agent policy gradients , COMA) 尝试通过联合训练智能体进行协同工作, 但需要对每个智能体的策略空间进行显式建模, 这在面对大规模智能体系统时可能会变得非常复杂[15].

## 2 基于学习机制的多智能体强化学习算法

从独立学习智能体到深度多智能体强化学习, MARL 在诸多研究领域都得到了广泛的应用, 但依然存在很多问题, 需要进一步的探索. 近年来, 元学习 (Meta-learning)、课程学习 (Curriculum learning, CL) 等新兴概念兴起, 在机器学习领域获得很多关注. 一些学者也将这类机器学习的机制应用于 MARL 中, 在提升 MARL 某些方面的性能上取得了惊人的效果. 例如将课程学习应用于 MARL 中, 引导智能体“学会学习”, 更快地学习到有效的策略. 将迁移学习应用到 MARL 方法中, 有效缓解真实任务场景中训练数据缺乏的问题等. 这些学习机制通过引入了额外的学习过程或者计算方法, 有效降低了 MARL 算法的计算开销和训练难度, 并在不同方面提升算法性能, 引导智能体学会了泛化性更强的策略. 借助学习机制提升 MARL 算法性能是未来的一个重要的方向, 相关的研究已经取得了不小的成绩, 并展现出巨大的发展潜力. 在本节中我们将讨论这一新兴的、活跃的、交叉领域的研究成果, 并探讨有待进一步探索的问题.

### 2.1 基于课程学习的多智能体强化学习

Bengio[16] 首先提出了课程学习的概念, 它是一种训练策略, 模仿人类的学习过程, 主张让模型先从容易的样本开始学习, 并逐渐进阶到复杂的样本和知识. 近年来, 有学者将课程学习引入多智能体强化学习中, 通过逐步增加任务难度和复杂性引导智能体学习, 这种渐进式的学习过程能够提高算法的训练速度及稳定性, 并帮助智能体更好地理解环境, 改善探索策略.

#### 2.1.1 缓解维度爆炸

课程学习可以从减少计算量的角度来缓解多

智能体数量增大而导致的存储和计算难度, 从而缓解梯度爆炸的问题, 减少训练的时间和硬件要求.

Zhou 等[17] 提出了一个基于种群的并行计算框架 MAlib, 核心是一个集中式任务调度模型, 该模型在自动课程策略组合上实现了训练任务的高度灵活性, 可以在不同的分布式计算范式上实现高效的代码重用和灵活部署. Zhao 等[18] 从小规模的场景开始训练, 逐步增加多智能体的数量来解决大规模的问题. 根据任务动态调整智能体数量不仅可加快学习速度、提升策略的学习效果, 并在训练过程中保持稳定性. Long 等[19] 设计了基于自注意力机制的网络架构, 该架构拥有固定的参数量并可以接受任意数量的智能体作为输入, 同时引入了进化选择过程, 在学习过程中选择更适应下一阶段的智能体. CM3 是是一种多目标多智能体合作算法, 基于课程学习将训练好的单智能体拓展到多智能体环境[20]. 上述工作的难点在于目前的多智能体强化学习算法无法有效处理动态智能体数量的情况. Wang 等[21] 设计了一种称为动态智能体数网络 (DyAN) 的新型网络结构来处理网络输入的动态大小, 并提出三种跨课程迁移机制来加速学习过程.

在多智能体强化学习环境中, 维度爆炸是一个很常见的挑战. 将课程学习引入多智能体强化学习中, 通过对强化学习任务分解、逐步加深神经网络、自适应调整等具体的方法, 有效降低了维度爆炸带来的计算问题, 从而提高多智能体强化学习算法的稳定性和可拓展性.

2.1.2 解决稀疏奖励

课程学习在稀疏奖励方面, 自动课程学习可用于向智能体提出辅助任务, 逐步引导其学习轨迹从简单任务到困难任务, 直到解决目标任务.

Chen 等[22] 提出一种变分自动课程学习方法, 由实体进展和两个任务扩展部分组成. 实体进展部分负责分阶段增加环境中的智能体数目, 任务拓展部分在智能体数量固定的前提下, 通过斯坦变分梯度下降方法将任务分布扩展到整个任务空间. Wang 等[23] 提出了技术人口课程 (Skilled population curriculum, SPC), 将自动课程学习与分层学习相结合, 通过逐渐增加任务难度与调整智能体数量进行策略学习. 在复杂状态动作空间方面, 逐渐增加智能体数量是一个有效学习的方法. Narvekar 等[24] 设计了一种基于自监督学习的评估机制, 在逐渐增加智能体的数量的过程中, 自动生成合适的课程.

课程学习在多智能体强化学习中可通过逐渐引入奖励丰富的任务、使用领域专家知识等方法, 协助智能体在训练初期获取更多的成功经验, 以克服稀疏奖励问题. 这个逐渐引入任务的过程可提高智能体适应稀疏奖励情境的能力.

2.1.3 设计课程引导智能体行为

课程学习可以在环境设计和训练过程中提升算法的性能, 例如在环境设置中, 可以逐步扩大智能体的视野; 在策略学习中可以逐步降低策略的复杂度, 帮助智能体掌握基本的技能.

Du 等[25] 设计了一种基于自我博弈的自动课程生成框架, 该框架优化了两个离线的策略学习者 Alice、Bob. 他们各自拥有一个“友好”的目标生成器和一个“不友好”的目标生成器, “友好”的目标生成器倾向于提出更可行的目标, 它们的合作有助于搭建整体学习的架构. “不友好”的目标生成器倾向于提出更具挑战性的目标, 他们之间的竞争则有助于突破智能体能力的界限.

OpenAI 的研究者们[26] 通过多智能体相互竞争, 构建捉迷藏的场景, 发现智能体自动创建了包含多回合不同策略的自动课程, 随着环境复杂性的提高而更好地拓展, 智能体表现出人类相关技能的行为. 如图 1 所示, 智能体在训练过程中学习到了六种独特策略, 每种都能帮助他们进入游戏的下一个阶段.

Yang 等[27] 强调课程设计应诱导智能体行为多样性以帮助发现更好的策略, 以及在遇到未知情况时能够保证鲁棒性. 特别在多智能体环境中, 应该在联合策略空间中定义课程任务的多样性, 同时考虑所有智能体的现有策略.

课程学习和多智能体强化学习结合有助于加速训练过程、提高稳定性、应对复杂性和提高泛化性能. 然而, 该类方法的局限性也非常明显, 设计一个有效的课程学习计划通常需要专业的知识和专家经验, 以确定课程增加的方式和顺序, 可能不适用于所有强化学习任务. 此外, 该方法侧重于逐步增加任务难度, 因此难以处理复杂的情景, 某些任务可能需要在没有明确课程的情况下进行训练. 表 1 总结了课程学习在多智能体强化学习中的应用.

### 2.2 基于演化博弈的多智能体分层强化学习

演化博弈论曾成功地解释了生物进化中的某些现象, 最为经典的早期工作是 1973 年 Smith 和 Price[28] 将其用来解释动物的斗争行为, 同时提出了演化稳定策略. Bloembergen[10] 研究了 MARL 和演化博弈论之间的关系, Hilbe 等[29] 的研究表明,

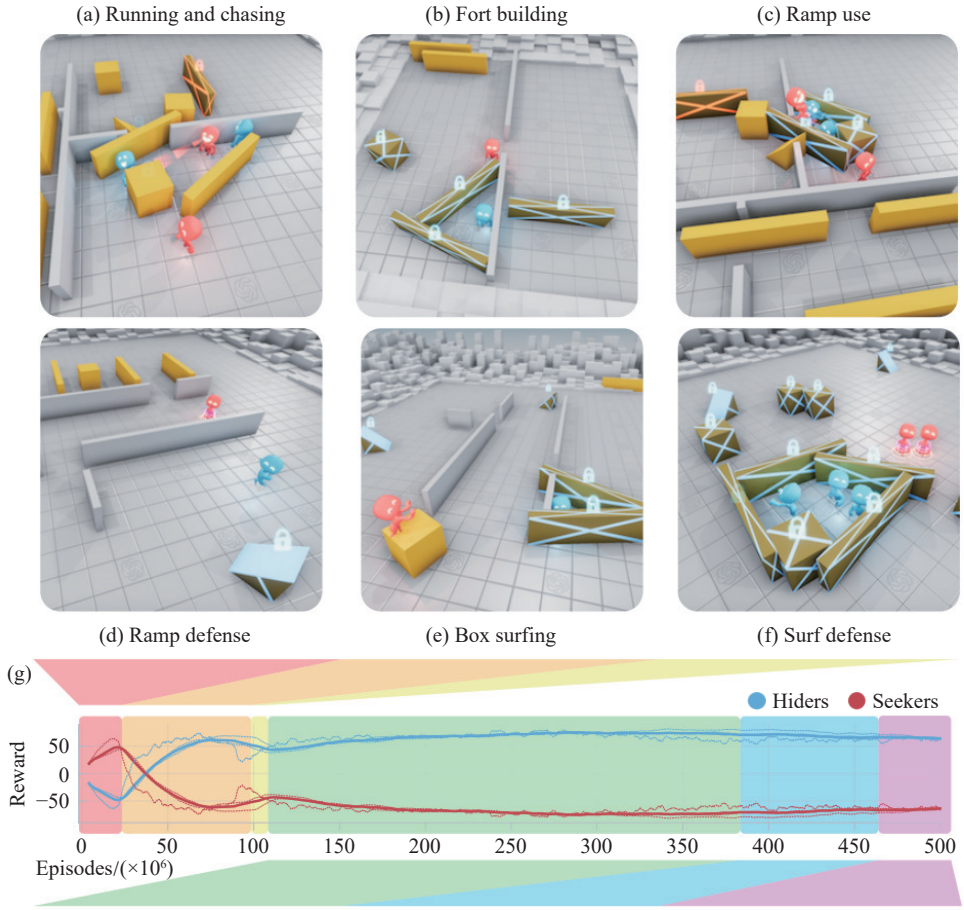


**图 1** 技能生成的训练过程. 通过奖励信号(如 $y$ 轴所示)在捉迷藏游戏中, 智能体会经历 6 个不同的出现阶段. (a) 探索者(红色)学会追逐躲藏者, 躲藏者学会粗暴地逃跑; (b) 隐藏者(蓝色)学习基本工具使用, 使用盒子, 有时甚至是现有的墙壁来建造堡垒; (c) 探索者学会使用坡道跳进躲藏者的庇护所; (d) 躲藏者很快学会将坡道移动到游戏区域的边缘, 很远他们将从那里建造堡垒并将其锁定到位; (e) 探索者得知他们可以跳跃从上锁的坡道到解锁的箱子, 然后从箱子上冲浪到躲藏者的避难所, 这是可能的因为环境允许智能体与盒子一起移动, 无论它们是否在地面上; (f) 隐藏者学会在建造堡垒之前锁上所有未使用的盒子; (g)3 次独立训练运行的平均值, 其中每个种子均以虚线显示[26]

**Fig.1** Training process for skill generation. Through the reward signal of hide-and-seek (shown on the $y$-axis), agents go through 6 distinct stages of emergence: (a) seekers (red) learn to chase hiders, and hiders learn to crudely run away; (b) hiders (blue) learn basic tool use, using boxes and sometimes existing walls to construct forts; (c) seekers learn to use ramps to jump into the hiders' shelter; (d) hiders quickly learn to move ramps to the edge of the play area, far from where they will build their fort, and lock them in place; (e) seekers learn that they can jump from locked ramps to unlocked boxes and then surf the box to the hiders' shelter, which is possible because the environment allows agents to move together with the box regardless of whether they are on the ground or not; (f) hiders learn to lock all the unused boxes before constructing their fort; (g) 3 independent training runs with each seed shown with a dotted line[26]

**表 1** 课程学习对强化学习算法的改进

**Table 1** Curriculum learning improves reinforcement learning algorithms

| Citations | Curricula generation | Main goal | RL algorithm |
|---|---|---|---|
| Zhao et al.[18] | Manual | Acceleration | IPPO |
| Long et al.[19] | Manual | Stability | MADDPG |
| Wang et al.[21] | Manual | Acceleration | DQN |
| Yang et al.[20] | Manual | Acceleration | Actor-critic |
| Narvekar et al.[24] | Automatic | Acceleration | Recursive Monte-Carlo |
| Chen et al.[22] | Automatic | Reward shaping | VACL |
| Wang et al.[23] | Automatic | Reward shaping | IPPO |
| Du et al.[25] | Automatic | Curriculum design | SAC |

演化博弈和随机博弈中重复交互的进化合作机制中, 互惠和环境回报反馈可以极大地增强合作倾向. Bloembergen 等[10] 利用演化博弈学习方法分析了各类多智能体强化学习方法的博弈动态, 并揭示了演化博弈论和多智能体化学习方法之间的深刻联系.

演化与强化学习结合的方法不通过估计值函数来推导最优策略, 而是使用非线性优化来直接搜索智能体的动作空间, 提升训练速度. 此外, 演化强化学习不完全依赖梯度下降法, 可以有效避免因梯度步骤而恶化的策略, 以更快地找到更优的策略.

2.2.1　优化神经网络

在深度学习领域, 往往使用梯度下降法来训练几千层或者几百层的神经网络, Uber 的研究者发现进化算法同样可以高效地为强化学习训练深度神经网络[30].

Moriarty[31] 早期的一项研究阐明了演化算法如何处理强化学习问题, 并在 2015 年由 Bloemberge 的研究证实[10]. 演化算法的高度并行性以及神经演化中的各种进化策略, 可以用来优化多智能体系统中的深度神经网络, 尤其在深度强化学习任务中可以取得有竞争力的表现. Khadka 等[32] 证实基因算法不需要反向传播, 已与流行的基于价值和基于梯度算法进行了比较, 算法在许多游戏里取得比现代深度强化学习更好的表现, 简单的基因算法可以训练带有超过 400 万参数的卷积网络, 同时由于算法有更强的并行能力, 可以运行得比常见方法更快. OpenAI 的研究者发现, 神经演化中的各种进化策略可以用来优化深度神经网络, 尤其在强化学习任务中取得具有竞争力的表现[33]. 实验表明, 只要提供了足够的计算资源, 进化策略近似计算的梯度在 MINST 数据集上达到 99% 的准确率, 此时, 近似梯度的进化策略就比梯度下降法更具优势.

演化算法作为一种无梯度方法, 适用于高度非线性的神经网络和高维度的参数空间, 在一些强化学习任务上表现良好. 旧算法和现代的海量计算的结合带来令人惊讶的结果, 其中的很多技巧可以用在深度神经网络尺度上, 给改进强化学习任务带来了很多启发. 但是, 演化学习与强化学习的结合也存在着一些挑战, 如需要精心设计种群规模大小, 变异强度等参数. 此外演化搜索策略的计算量也大大增加, 需要同时在上百个甚至上千个 CPU 集群上运行, 对计算性能提出了更高的要求.

2.2.2　平衡探索与挖掘

在多智能体的环境中, 智能体在彼此交互的过程中会周期性地调整自己的策略. 而演化思路正是一种可以高效模拟这些交互的方法, 从而改进了策略的探索.

Gomes 等[34] 证明演化算法可以进化出具有不同策略的智能体, 它可以把进化策略的优化能力和可拓展性与神经演化中独有的方法结合起来, 用一个鼓励各自做出不同行为的智能体群落提升强化学习任务中的探索能力. Deepmind 的 Jaderberg 等[35] 提出了 PBT 算法, 他们发现了一个超参数设置的时间表, 而不是遵循通常的次优策略, 即试图找到一个单一的固定设置用于整个训练过程. 只需对典型的分布式超参数训练框架稍加修改, 该方法就可以对模型进行鲁棒和可靠的训练, 并且 Jaderberg 等[12] 和 Vinyals 等[36] 分别在《夺旗争霸》和《星际争霸》中证实, 基于人群的训练 (Population based training, PBT) 有着实现超越人类的行为的强大力量. 如图 2 所示, 基于群体的神经网络训练初始时就像随机搜索一样, 并行执行很多个不同超参的任务, 从而创建了一个神经网络的“群体”, PBT 会周期性地将表现不好的模型替换掉. Conti 等[37] 研究发现, 新颖搜索算法 (Novelty search) 不仅在基因算法的效果基础上得到提升, 甚至还可以处理反馈函数带有欺骗性、奖励函数稀疏的情况, 还可以保持同等的可拓展性. Liu 等[38] 证明进化策略在非平稳性和部分可观的情况下取得很好的性能, 因为它不断地使用和发展一群智能体, 而不是单一智能体, 优化的是数个解决方案的最优分布, 而不是单独一个最优解决方案, 因此在受到足够多的参数扰动时, 进化策略比梯度下降法训练的模型具有更好的鲁棒性.

相关的理论和实验已经证明了演化等方法可以取代梯度下降等现有主流方法用来训练深度强化学习模型, 同时取得更好的表现. 采用进化算法的知识对多智能体深度强化学习算法改进, 将鼓励智能体对具有欺骗性稀疏奖励的任务进行探索, 有效地促进智能体之间的协作行为, 在多智能体强化学习领域中开发出新的解决方案.

### 2.3　基于元学习的多智能体强化学习

元学习通过发现并推广不同任务之间的普适规律来解决未知难题, 可以泛化到差异很大的新领域中, 提升在强化学习任务上的表现. 元强化学习是强化学习范畴内的一个方法, 同时也是元学

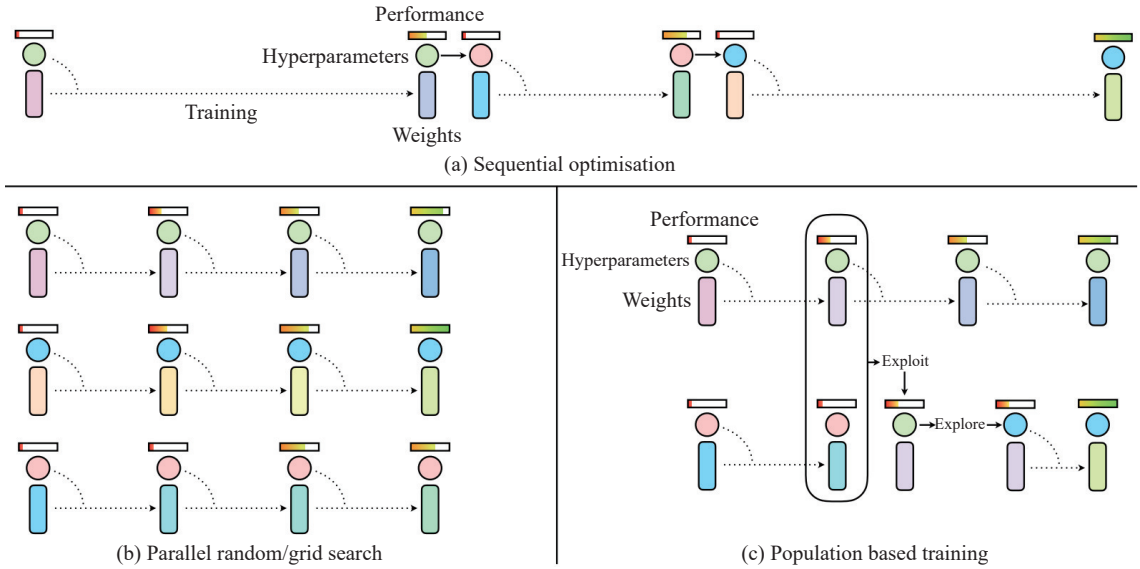


**图 2** PBT 的演化过程. (a) 顺序优化需要多次训练运行完成(可能提前停止), 之后选择新的超参数并建立模型使用新的超参数从头开始重新训练.这是一个本质上连续的过程, 并导致很长的时间超参数优化时间, 但使用最少的计算资源; (b) 并行随机/网格搜索超参数使用不同的权重初始化和超参数并行训练多个模型, 其中认为其中一个模型将得到最佳优化. 这仅需要运行一次训练的时间, 但需要使用更多的计算资源来并行训练许多模型; (c) 基于群体的训练像并行搜索一样开始, 随机采样超参数和权重初始化. 但是, 每次训练运行都会异步评估其性能定期. 如果群体中的模型表现不佳, 它将利用其余群体通过用性能更好的模型替换自身, 并且它将通过修改更好的模型来探索新的超参数在继续训练之前, 调整模型的超参数. 这个过程允许在线优化超参数, 并且将计算资源集中在最有可能产生的超参数和权重空间上良好的结果. 结果是一种超参数调整方法, 虽然非常简单, 但学习速度更快, 成本更低计算资源, 通常还有更好的解决方案[35]

**Fig.2** Evolution of PBT: (a) sequential optimisation requires multiple training runs to be completed (potentially with early stopping), after which new hyperparameters are selected and the model is retrained from scratch with the new hyperparameters. This is an inherently sequential process and leads to long hyperparameter optimisation times, though uses minimal computational resources; (b) parallel random/grid search of hyperparameters trains multiple models in parallel with different weight initialisations and hyperparameters, with the view that one of the models will be optimised the best. This only requires the time for one training run, but requires the use of more computational resources to train many models in parallel; (c) population based training starts like parallel search, randomly sampling hyperparameters and weight initialisations. However, each training run asynchronously evaluates its performance periodically. If a model in the population is under-performing, it will exploit the rest of the population by replacing itself with a better performing model, and it will explore new hyperparameters by modifying the better model's hyperparameters before training is continued. This process allows hyperparameters to be optimized online, and the computational resources to be focused on the hyperparameter and weight space that has the most chance of producing good results. The result is a hyperparameter tuning method that while very simple, results in faster learning, lower computational resources, and often better solutions[35]

习的一个特例, 其主要目标是快速适应新的任务. 与传统的多智能体强化学习算法相比, 多智能体元强化学习在较少的样本上实现了更高的性能, 并生成通用性更强的策略. 假设是多智能体元学习任务对服从同一分布 $p(T)$, 学习一个参数为 $\theta$ 的策略函数, 该函数能够最小化特定任务 $T$ 的损失函数 $\mathcal{L}s$. 因此我们目标便成了学习一个过程 $\theta' = u_\psi(D_T^{\text{tr}}, \theta)$, 可以使用很小的数据集 $D_T^{\text{tr}}$ 快速适应新任务 $D_T^{\text{test}}$:

$$\min_{\theta,\psi} \mathbb{E}_{T\sim p(T)}\left[\mathcal{L}s\left(D_T^{\text{test}}, \theta'\right)\right], \text{ s.t.}\theta' = u_\psi\left(D_T^{\text{tr}}, \theta\right) \quad (2)$$

不同的元强化学习算法的区别在于学习过程的具体实施形式. 当前, 基于元学习的多智能体博弈方法主要分为两大类, 第一类是基于上下文 (Context), 第二类是适应学习 (Learning to adapt).

2.3.1 基于上下文的多智能体元学习方法

基于上下文的多智能体元学习方法使用 Context 来估计隐藏的任务嵌入来实现快速适应[39], 即根据任务的历史样本推断该任务的信息, 并将其整合到新定义的隐层变量中, 然后学习算法就可根据所得的隐层变量调整策略. 因此, 此类算法的目标不仅有学习出如何对任务进行推断, 还有学习如何根据任务推断的结果优化策略.

谷歌团队[40] 提出了一个基于 Transformer 的元学习分布框架 (General-purpose in-context learning, GPICL), 将 Transformer 和其他黑盒模型被元训练成通用的上下文学习器, 在小样本上进行训练学习并推广到没有训练过的任务中, 同时在深度数据表示上学习线性模型来解决非线性回归任务.

现有的研究通常假设元学习的任务状态分布是固定的[41−46],这种框架同样适用于多智能体博弈学习环境. Vezhnevets 等[47] 针对多智能体博弈中存

在的泛化问题, 提出具有隐藏信息和复杂非传递奖励函数的多智能体博弈方法, 经过环境测试表明, 该方法具备出色的泛化能力.

基于上下文的多智能体元学习方法的目标不仅学习出如何对任务进行推断, 还要学习如何根据任务推断的结果优化策略. 这种方法使得多智能体强化学习算法具有更高的泛化能力, 在不同的环境中快速适应新的任务, 更好地适应和协同工作.

2.3.2 基于自适应的多智能体元学习方法

基于自适应的多智能体元学习方法为基础, 所学习的学习过程$\theta' = u_\psi(D_T^{\mathrm{tr}},\theta)$为策略梯度上升的过程, 在此类算法中, 目标函数经过梯度下降优化后, 在解决新任务时可以仅通过少数步骤的梯度变换实现快速适应新任务, 提高了算法的通用性.

Finn 等[48] 首先提出了一种基于策略参数更新的元学习算法(Model-agnostic meta-learning, MAML). MAML 在梯度下降过程中找到对任务敏感的模型参数, 并学习模型中的初始化参数, 使得该套参数初始化的模型仅通过少量样本上的一次或几次梯度更新就能够最大化新任务的性能. 假设模型为$f$, $\alpha$和$\beta$为学习步长, MAML 算法采样任务, $T_i \sim p(T)$, 则梯度下降更新参数$\theta$:

$$\theta_i' = \theta - \alpha\nabla_\theta L_{T_i}(f_\theta) \tag{3}$$

当采样任务全部完成之后, 利用训练过程中的采样轨迹和计算所得损失进一步更新参数$\theta$:

$$\theta \leftarrow \theta - \beta\nabla_\theta \sum_{T_i\sim p(T)} L_{T_i}(f_{\theta'}) \tag{4}$$

MAML 算法是基于梯度的元学习算法的基础, 基于适应的多智能体元学习方法[49] 都是在此基础上改进的. Wu 等[50] 利用元学习方法同时可以生成难被利用和多样性对手, 引导智能体自身的策略迭代提升. Al-Shedivat 等[51] 针对多智能体博弈对抗, 提出基于梯度的元学习算法, 使得智能体可以在非平稳环境中具备连续的适应能力.

Foerster 等[52] 提出了具有对手学习意识(Opponent-learning awareness, LOLA)的元学习方法, 该方法通过在策略梯度计算时引入额外的修正项, 来考量每个智能体策略对环境中其他智能体学习过程的影响.

Kim 等[49] 利用元学习和对手建模, 设计了第一个元多智能体策略梯度定理(Meta-multiagent policy gradient theorem, Meta-MAPG), 考虑智能体自身的非平稳策略动态和环境中其他智能体的非静态策略动态.

基于自适应的多智能体元学习方法可以根据每个智能体的个体特征和学习历史来自适应调整学习策略. 这意味着不同的智能体以不同的方式适应环境, 更好地满足其特定需求, 从而提高了算法的灵活性和多模态性. 多智能体元学习是一个相对新兴的研究领域, 存在着明显的弊端, 如目前的理论基础相对不足, 还需要更多的理论研究来支持这些方法的发展. 其次, 将元学习方法与多智能体系统结合时, 可能会遇到迁移问题, 即一个任务中学到的知识和策略如何有效地迁移到另一个任务中, 这需要更多的相关研究.

### 2.4 基于分层学习的多智能体强化学习

多智能体分层强化学习使用分层策略将 MARL 问题抽象成不同的层级, 能够有效解决维度爆炸的问题. 算法通过将长期 MARL 问题分解为子问题或子任务, 以便高级别的策略通过选择最佳子任务作为更高级别的操作, 而子任务本身可能是一个强化学习问题, 需要通过较低级别的策略学习来解决它. 目前, 多智能体分层强化学习主要可分为基于选项 (Option) 和基于目标 (Goal) 两大类.

2.4.1 基于选项的多智能体分层强化学习

基于选项的多智能体分层强化学习采用 Option 分层强化学习方法对学习任务进行时序抽象, Option 本质上为在某状态子空间里完成相应子任务的动作序列, 而 Option 也作为一种特殊的动作跟基本动作一起构成动作集, 通过上下 Option 之间相互调用形成分层结构. 除此之外, Option 需要通过先验经验或知识来获得[53].

Tang 等[54] 则针对 StarCraft 问题, 根据作战规则不同, 作战要素和动作空间组合不同设计 101 种输入特征向量选项, 实现了订单生产的强化学习. Zhang 等[55] 针对 RTS 游戏领域中的需要宏观策略和微观策略才能获得满意的性能, 提出了一个分层框架, 其中智能体通过模仿学习执行宏观策略, 并通过强化学习进行微观操作. Liang 等[56] 在自动驾驶车辆协调问题中将低层次的个体控制归结为单智能体强化学习; 在高层离散动作空间中建立对手建模网络, 在学习过程中对其他智能体的策略进行建模, 有效提高了多智能体的协作性. Bacon 等[57] 提出了一种目标-执行者 (Option-critic) 学习方法, Option-critic 在 AC 结构的基础上增加了对 Option 的切换和执行, 其通过深度神经网络来寻找任务之间的边界, 可以自动地学习出 Option

的策略和 Option 的切换函数. 图 3 为 Option-critic 的架构图.

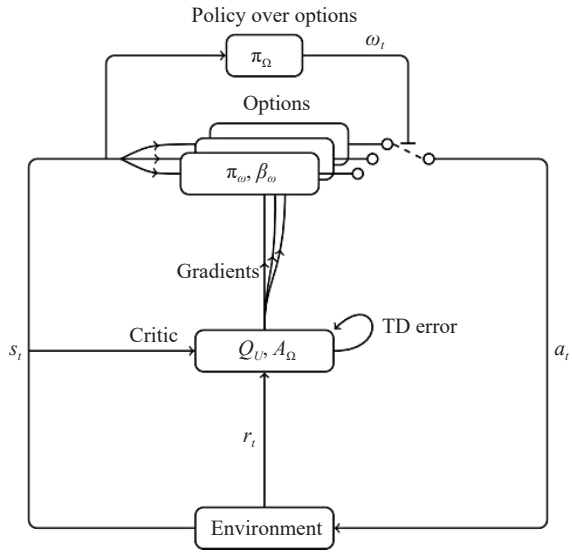


**图 3** Option-critic 架构图[57]

**Fig.3** Option-critic architecture diagram[57]

基于选项的多智能体分层强化学习的可执行时态拓展动作的特点对强化学习摆动期的缩短和效率提高均有一定的促进作用, 但需要基于先验知识确定 Option, 所以基于选项的分层强化学习在未知环境中的适用性还有待提高[58−60].

2.4.2 基于目标的多智能体分层强化学习

基于目标的分层强化学习中, 上层的控制器需要在较长的时间跨度上确定一个目标, 而下层的控制器则在较短的时间跨度上选择动作来达到这一目标[61−62]. 不同的文章对于目标的定义也是不同, 有些是根据特定的场景进行定义, 有些场景的任务复杂难分, 则通过端到端的方式进行自动任务生成并分层[63].

Kulkarni 等[64] 提出了一种分层 $Q$ 值方法, 其通过构造两个层级的算法, 顶层用于决策, 确定下一步的目标, 底层用于具体行动决策, 在 Montezuma's Revenge 游戏中取得了较好的效果. Xu 等[65] 提出了一种面向目标的多智能体分层协作框架 (Hierarchical target-oriented multi-agent coordination, HiT-MAC) 解决多个传感器目标覆盖的问题. 该框架主要分两层, 包括一个集中式的协调者和多个分布式的执行者, 协调者每 $K$ 步收集各个执行者的观测信息, 进行全局规划, 为每个执行者分配特定的任务目标, 执行者每步通过采取一系列基本动作来完成指定任务, 图 4 为 HiT-MAC 框架.

Ma 等[66] 将封建强化学习 (Feudal reinforcement learning, Feudal RL) 和多智能体 A2C 算法 (Multi-agent A2C, MA2C) 结合形成 FMA2C 方法, 根据目标设计构建了一个分层设计的多智能体学习结构, 每个智能体都有自己的上层策略和共享的下层网络, 首先训练低层次的网络进行简单的运动控制, 以某个潜变量或者目标 Goal 作为条件而上层策略需要学会向下层目标提供目标.

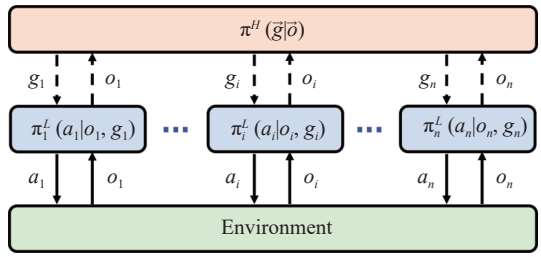


**图 4** HiT-MAC 框架[65]

**Fig.4** HiT-MAC framework[65]

Vezhnevets 等[67] 提出的封建网络算法 (Feudal networks, FuNs), 通过在多个层次上解耦端到端学习, 即允许它使用不同的时间分辨率, 并且使用了 Manager 模块和 Worker 模块. Manage 设定抽象的目标, 训练目标是尽可能输出隐层特征空间中有利的转移方向, Worker 在环境的每个时间步产生原始动作, 尽可能达成目标, FuNs 在涉及长期信用分配或记忆的任务上表现优异.

基于目标的分层强化学习中, 上层控制器的训练准则是让输出目标尽可能满足所设定的语义信息, 即让目标尽可能符合较好的状态转移方向, 而下层控制器的训练准则是让环境状态转移尽可能地满足目标, 从而使得智能体能够更加高效地学习到最优策略[56].

目前, 多智能体分层强化学习已成为协同决策和解决复杂任务的有利工具, 但仍有很多固有的挑战, 例如, 如何分解任务, 确定不同层次的目标, 以及设计有效的分层策略. 在多智能体系统中, 不同智能体的决策可能会相互影响, 因此, 需要确保不同层次的策略之间能够有效协同和通信. 分层强化学习需要在不同层次上进行有效的探索, 以找到最佳策略. 表 2 总结了我们调研的分层强化学习算法.

### 2.5 基于迁移学习的多智能体强化学习

迁移学习是一种机器学习领域的方法[68−70], 它利用在其他领域上学到的外部知识来改善在目标任务上的性能. 虽然迁移学习技术已在监督学习领域得到广泛研究[71], 但它仍然是强化学习的一个新兴主题[72]. 为降低多智能体强化学习算法的复杂性, 迁移学习重用来自先前经验或其他智能

**表 2**　分层强化学习的算法总结

**Table 2**　Summary of algorithms for hierarchical reinforcement learning

| Citation | Classify | Algorithm | Baseline |
|---|---|---|---|
| Tang et al.[54] | Option | NNFQ/CNNFQ | NN/CNN |
| Zhang et al.[55] | Option | HRL | PPO |
| Liang et al.[56] | Option | MAHRL | Actor-critic |
| Bacon et al.[57] | Option | Option-critic | DQN |
| Xu et al.[65] | Goal | HIT-MAC | Actor-critic |
| Liu et al.[59] | Goal | FMA2C | MARL |
| Vezhnevets et al.[67] | Goal | FuNs | DQN/Option-critic |

体的知识的特点拥有巨大潜力.

之前的研究[73]将从智能体中得到的迁移知识分为智能体内迁移(即在新任务中重用之前智能体获得的知识)和智能体间迁移(解决何时以及如何传递知识), 这样分类是不严格的, 二者之间的界线比较模糊. 本文中我们将智能体知识迁移分类为同构智能体知识迁移与异构智能体知识迁移, 两者的区别在于相互通信知识的智能体之间是否具有相同的模型结构.

2.5.1　先验知识迁移

虽然强化学习算法在很大程度上是模仿人类自主学习的过程, 但实际的强化学习算法仍与人类的学习过程有很大不同[9]. 在多智能体迁移学习中, 自然会想到将人类专家的先验知识迁移给智能体[74]. 例如, 专家可能能够设计信息丰富的奖励塑造或规则, 进而加速智能体的训练过程. 如何正确高效利用专家的知识并非易事, 在本小节将讨论当前在这方面的主要方法.

MacGlashan 等[75]提出了依赖专家反馈收敛 AC 结构, 专家反馈某一动作对当前策略的影响, 使得网络从反馈中学习并收敛到局部最优. 这项工作在模拟的网格世界任务与现实机器人任务中都展示出良好的表现. Li 等[76]在无人机自主空战场景中, 使用基于专家演员的软演员-评论家算法(Expert actor-based soft actor critic algorithm), 该算法利用专家先验知识构建经验重播缓冲区, 利用少量专家经验增加样本的多样性, 从而使智能体在算法陷入局部时能够扩展探索, 提高智能体的探索效率, 如图 5 所示.

DeepMind 团队 Abramson 等[77]提出了使用人类反馈强化学习来改进智能体的学习能力, 首先收集人类在模拟 3D 世界中与智能体交互的数据, 之后要求注释专家记录他们认为智能体朝着人类指示的目标移动的时刻, 最后通过这些记录数据构建相应的奖励模型. 该研究成功利用人类专家的经验知识改善了智能体的能力.

先验知识迁移的方法可以解决样本效率低和训练时间长的问题, 加速了智能体在新任务中的学习过程. 然而迁移过程需要谨慎设计, 确保先验知识在新任务中是适用的, 如果任务目标差异过大, 可能会产生负面的影响. 同时过度依赖先验知识可能限制了智能体自主学习的能力, 需要平衡先验知识和环境交互学习的重要性.

2.5.2　同构智能体知识迁移

同构智能体间的知识迁移是指在模型结构相

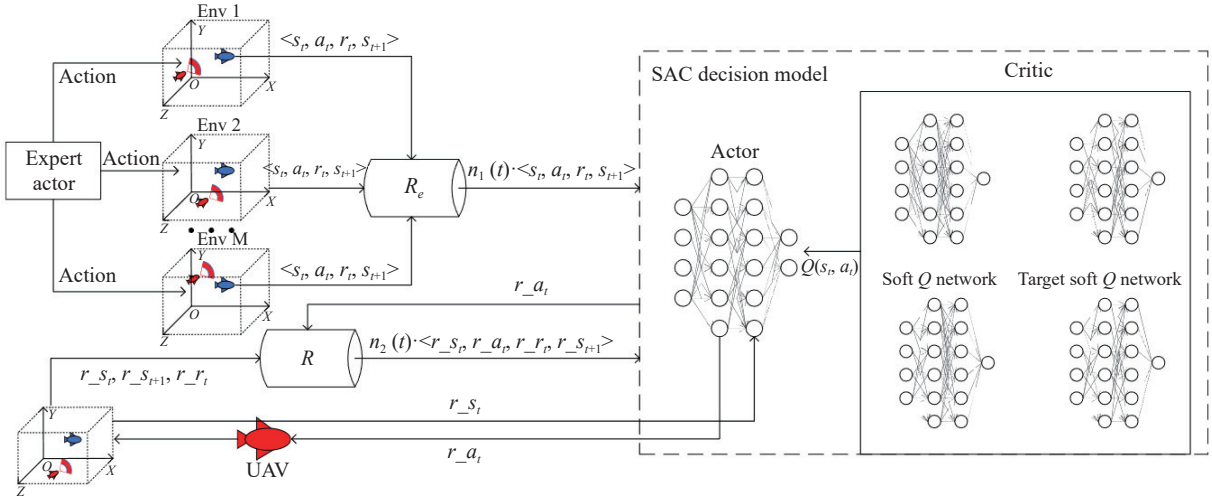


**图 5**　基于专家的软演员–评论家算法示意图. 无人机使用专家 Actor 根据状态 $s_t$ 做出决定, 获得动作 $a_t$, 并执行动作以获得奖励和新状态 $s_{t+1}$, 直到成功摧毁敌人. 在不同的空战环境中重复上述过程, 并将获得的专家经验样本 $(s_t, a_t, r_t, s_{t+1})$ 存储在专家重放缓冲区[76]

**Fig.5**　Schematic of Expert-based soft actor-critic algorithm. The drone uses an expert actor to make decisions based on state $s_t$, obtain an action $a_t$, and perform actions to obtain a reward $r_t$ and a new state $s_{t+1}$ until the enemy is successfully destroyed. Repeat the above process in different air combat environments, and store the obtained expert experience samples $(s_t, a_t, r_t, s_{t+1})$ in the expert replay buffer[76]

同或者参数设置相同的智能体之间共享和传递信息, 从而加速彼此的训练过程. 同构的智能体间的通信是直接的, 无需做过多处理. 通过知识迁移, 同构智能体可以更快地学习新任务, 提高性能并且更好地适应变化的环境.

Liu 等[78] 通过智能体间的知识迁移, 解决了多智能体系统中, 因智能体数量逐渐增多而训练速度过慢的问题. 设置已接受训练的教师智能体, 将学生智能体的观察结果作为输入, 并输出建议的动作和值作为监督信息来指导学生智能体. Hu 等[79] 将参数作为知识进行迁移共享, 以解决多智能体网络中, 通信维度指数增长的大规模优化问题. 该研究所提出的参数共享迁移学习方法可以加速多智能体强化学习算法的收敛速度, 同时使用了相关对齐方法, 在共享参数的同时保持了局部经验特征. Liang 等[80] 将联邦学习与迁移学习结合起来, 实现在不同环境中的同构智能体进行异步学习与知识聚合, 将模拟器上的预训练与实际场景中的微调结合. Li 等[81] 首次使用模糊逻辑实现策略到行为的映射, 大大降低了同构多智能体强化学习的成本. 提出了模糊智能体的概念, 打破了智能体与策略的一一对应关系. 通过使用模糊智能体与代替多智能体与环境交互, 并隐式保留了智能体间的耦合关系, 以较低的计算复杂度, 实现大规模同构智能体系统的轻量化学习.

同构型的智能体拥有如行为、目标和领域知识等较多的共性特点, 可以通过知识迁移的方式提升学习的效率和速率, 同构智能体可以通过模型参数共享、经验共享等方法, 快速从其他智能体的经验中受益, 提高学习效率.

2.5.3 异构智能体知识迁移

异构智能体的知识迁移是富有挑战的, 因为不同结构的智能体在学习和表示能力上存在差异, 这导致他们之间难以进行有效的知识利用, 需要寻找有效的方法来将不同模型之间的知识转化和对齐[82−83]; 此外, 异构智能体的知识迁移可以帮助智能体在一个领域中学到的知识迁移到另一个不同的领域中, 从而实现跨领域的技术应用和创新.

Zhou 等[84] 提出一个基于稀疏交互的多智能体强化学习算法, 通过局部的 $Q$ 值迁移来提升智能体的学习速度, 解决多智能体系统中的学习与协同问题. 与传统多智能体稀疏交互算法相比, 该研究引入均衡的概念, 使得智能体可以很容易找到接近最优的策略. Li 等[85] 将迁移学习与知识蒸馏结合起来, 实现异构智能体的联邦学习. 迁移学习是多智能体合作的前置, 知识蒸馏作为多智能体间通信的基础. Shi 等[86] 则提出了一种在异构智能体间跨任务迁移知识的方法. 该方法受到渐进神经网络启发, 将策略的特征作为知识进行迁移, 并引入注意力模块增强传输效果. 智能体之间通过横向连接接受特征知识, 并且有效避免了在源任务与目标任务完全不同情况下的负迁移. 如图 6 所示.

不同于同构智能体, 当训练环境是由大量异构智能体构成时, 需要解决异构智能体的信用分配、过估计、可拓展性等问题, 同时需要考虑到异构智能体间特征与知识的对齐及转化. 此外, 因为异构系统的复杂性更高, 需要设计更为灵活的迁移学习算法和适应性策略.

总体而言, 将迁移学习应用于 MARL 中, 为复杂多智能体系统带来了显著的优势. 迁移学习通过特征提取和共享、策略迁移等方法, 加速了智能体的学习速率, 使得多智能体可以尽快适应不同的任务, 提高了模型的泛化能力. 然而, 这一结合的方法也存在着诸多难题, 如迁移学习可能导致负迁移, 即在新环境中的性能下降. 这可能是因为源任务和目标任务之间的差异过大, 使得迁移的知识反而不适用于新任务. 此外, 迁移学习和多智能体强化学习的结合可能增加计算复杂性, 尤其是在异构多智能体系统中, 需要有效地管理不同智能体之间的知识迁移.

表 3 展示了我们调查工作的具体情况.

## 3 基于学习机制的 MARL 实际应用

如何在多智能体强化学习算法充分发挥学习机制的优势, 将 MARL 和学习机制结合起来进行优势互补, 解决实质性的问题是研究基于学习机制的 MARL 的重要意义所在. 这一融合算法已经在多个实际领域得到了应用, 如博弈编队控制、交通控制、机器人协同、智能游戏、经济学和社会学领域等.

编队控制和交通控制是 MARL 的一个重要应用, 其目的是使智能体找到一条从起点到终点的最优路径, 在这个过程中还需要完成避障、搜集问题以及导航到多个目标的任务. Liang 等[56] 采用基于分层的 MARL 思想, 在自动驾驶车辆协调问题中将低层次的个体控制归结为单智能体强化学习; 在高层离散动作空间中建立对手建模网络, 在

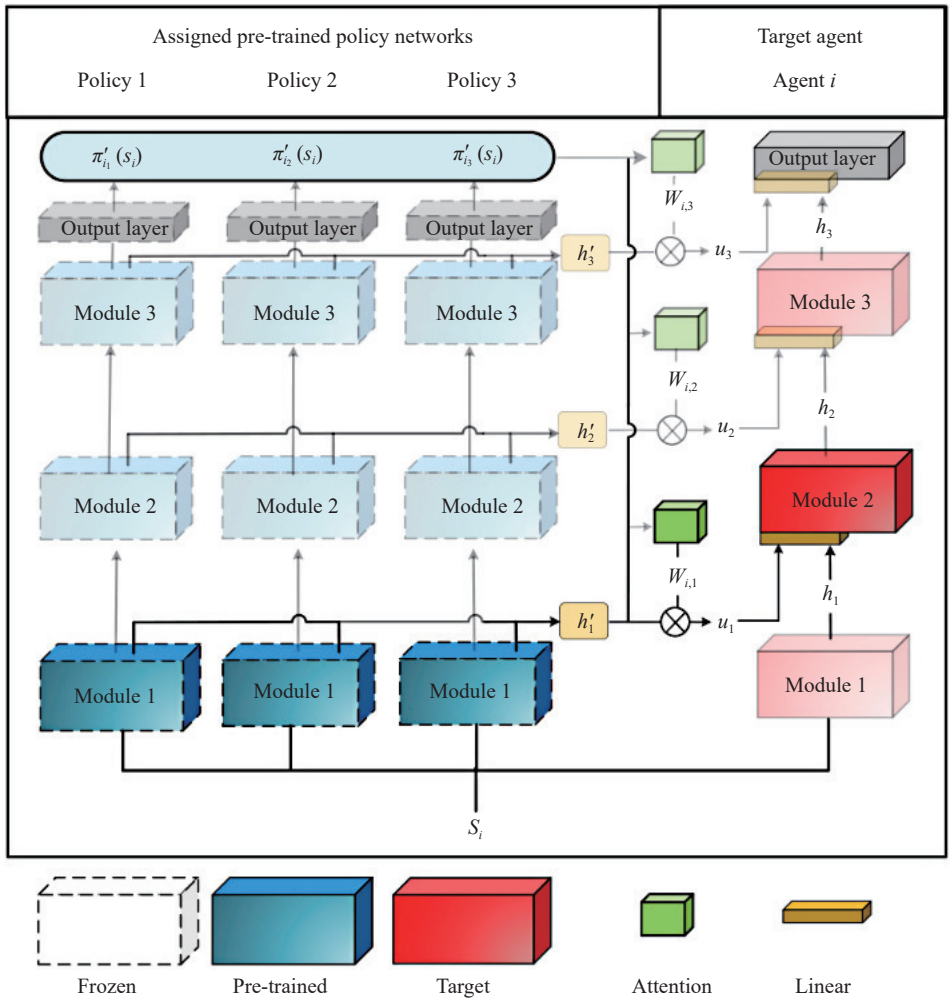


**图 6**　目标任务中智能体横向传输的示例, 将三个预训练的策略网络分配给智能体 $i$, 用于传输知识, 并且所有策略网络$\pi_i$中都有三个模块和一个输出层. 对每个智能体$i$, $h'_k$代表所有分配的策略在第$k$个模块的输出, $w_{i,k}$代表由注意力模块输出的第$k$个模块的权重, $u_k = h'_k \otimes w_{i,k}$[86]

**Fig.6**　Example of Agent $i$ lateral transfer in the target task, assigning three pretrained policy networks to agent $i$ for transferring knowledge, and there are three modules and one output layer in all policy networks $\pi_i$. For each agent $i$, $h'_k$ represents the output of all assigned strategies in the $k$-th module, $w_{i,k}$ represents the weight of the $k$-th module output by the attention module, $u_k = h'_k \otimes w_{i,k}$[86]

**表 3**　迁移强化学习算法总结

**Table 3**　Summary of Transfer Reinforcement Learning Algorithms

| Citations | Subject | Transferred knowledge | Algorithm |
|---|---|---|---|
| MacGlashan et al.[75] | Prior knowledge | Reward shaping | Actor-critic |
| Liu et al.[78] | Prior knowledge | Reply buffer | Actor-critic |
| Li et al.[76] | Homogeneous agent | Action & Value | MADDPG-INA |
| Hu et al.[79] | Homogeneous agent | Model parameters | ACER |
| Liang et al.[80] | Homogeneous agent | State & action | DDPG |
| Li et al.[81] | Homogeneous agent | Strategy | Fuzzy agent |
| Shi et al.[86] | Heterogeneous agent | Strategy | MADDPG |
| Li et al.[85] | Heterogeneous agent | Model parameters | Any |
| Zhou et al.[84] | Heterogeneous agent | Local Q function | Q-learning |

学习过程中对其他智能体的策略进行建模, 有效提高了多智能体的协作性. 其中, Lin 等[87] 参考了基于知识迁移的多智能体强化学习算法, 多种合作型算法应用在大型编队控制问题上, 他们的方法聚焦于如何平衡分配交通资源以提升交通效率, 减少拥堵, 该方法使用参数共享机制保证多个

车辆间的协同.

在机器人协作中涉及到智能体之间的交互, 需要考虑状态动作对的维度过大、智能体之间通信协作等问题. Liu 和 Tan[88] 对智能体的邻居关系建模, 使用注意力机制的编码器来聚合任意数量邻近智能体的特征, 同时采用参数共享的方式来降低计算量, 提高了算法的可拓展性, 可以用于大规模智能体训练. Yang 等[20] 将多智能体协作问题重新划分为两阶段课程, 引导多智能体探索个人目标的实现和然后再学会合作实现他人目标, 使用功能增强方案来桥接整个课程的价值和策略功能, 满足了有效探索, 获得了更高的性能、学习速度和鲁棒性. Xu 等[65] 构建了分层的多智能体强化学习算法 HiT-MAC, 使得上下层分别进行动态任务目标分配和目标追踪的任务, 应用于有向传感器网络中, 进行主动的目标跟踪, 获得了更高的鲁棒性和更大的覆盖率.

游戏领域一直是多智能体强化学习研究的重点, 也是其迈向更广阔领域的基石. 电子游戏提供了大量的交互样本, 满足训练 MARL 算法的需求. 游戏环境可以分为两类: 一类用来提升算法的通用性, 如 Atari 2600、MuJoCo; 另一类用来处理复杂的游戏场景, 如 AlphaStar 等. Uber AI 的研究人员使用演化强化学习, 这种新的探索方式可以提高多智能体强化学习在 MuJoCo 中许多任务中的表现, 有效避免具有欺骗性的局部极值[31]. OpenAI[23] 促进智能体自动创建多回合不同策略的自动课程, 随着环境复杂性的提高, 智能体逐渐学会不同的策略, 表现出了人类相关技能的行为, 该算法被证明了能引导智能体更好的在高维的策略空间中进行搜索. DeepMind[36] 在星际争霸等游戏中使用的 PBT, 这种方法基于超参数优化和遗传算法的思想, 动态调整超参数并可以进行并行计算, 有效增强了算法探索能力, 结果表明智能体之间学会了竞争或者协作的多种策略, 其游戏能力达到大师级水平.

基于学习机制的 MARL 算法也在博弈论研究和社会福利等社会问题中也得到了许多应用, 如 Schmid 等[89] 将经济学中的交易规则参考知识迁移的方式引入到多智能体系统中, 在该系统中, 智能体的动作、状态、回报等参数都被看成可以互相交易的资源. 目前, 基于学习机制的 MARL 已经在不同领域的应用取得了可观的成果, 有效突破了传统的 MARL 在这些应用领域中的瓶颈, 期待有更多的相关算法得到研究和推广.

## 4 基于学习机制的 MARL 前景展望

基于学习机制的多智能体强化学习作为一新兴概念, 起步时间较晚, 理论成熟性较低, 其发展潜力巨大, 前景相当可观. 本节接下来将对基于学习机制的 MARL 算法的前景展望进行简单的阐述.

### 4.1 更多的理论研究和基准工作

现有研究的实验结果表明基于学习机制的多智能体强化学习算法是有效的, 但很少有工作进行理论证明并设计标准的测评环境来评估它们. 在现有的文献中, 训练环境和指标在不同的应用中是多样化的, 不足以全面地评估多智能体强化学习算法的性能. 如目前构建基于课程学习的 MARL 中的课程通常基于人类的直觉和经验, 缺少对 MARL 领域课程设计进行合理的评估. 在演化强化学习的研究中, 研究人员发现在某些强化学习任务中, 采用基因算法取得了比梯度下降法更具竞争力的表现, 但这一发现缺少进一步的理论论证和研究, 因此这个方法更广泛的意义和结果目前仅限于研究人员的猜测中.

一些基于学习机制的 MARL 在常用的强化学习任务上进行评估, 如 OpenAI Gym 和 Multi-agent particle environment(MAPE), 另一些论文仅在特定的训练环境进行评估, 并对于如何在所用的训练环境上分割训练和测试任务缺少统一的标准. 设计并采用统一的基准, 将更好地评估基于学习机制的 MARL 算法的泛化性. 近年来, 涌现了许多基于学习机制的 MARL 算法, 但是却鲜少有人探讨在如何设计 MARL 环境可以作为这类算法的标准测试环境.

### 4.2 面向大规模多智能体应用研究

目前的 MARL 算法在大规模博弈对抗场景中存在着维度爆炸的问题, 极大地限制了算法的可扩展性, 而当前相关研究正在从简单小规模博弈向大规模复杂任务及通用博弈场景聚焦, 融合学习机制的相关算法为 MARL 在更广泛的研究场景中提供了参考.

Ghosh 等[90] 将强化学习中的泛化问题定义为认知部分观测马尔可夫决策过程(Epistemic POMDPs), 为多智能体博弈决策泛化问题的求解提供了解决思路. Yang 等[91] 根据平均场思想提出的平均场 $Q$ 学习和平均场 Actor-critic 方法, 为解决大规模智能体学习问题提供了参考. Huang 等[92] 针对多智能体的策略学习问题, 提出了一种模块化共享策略 SMP(Shared modular policies),实现用同一策略

控制不同的智能体, 极大提高了智能体策略的泛化性. 数据知识双驱动的智能体策略学习方法也受到了广泛的关注, 基于领域知识的智能体策略学习具有较强的可解释性[93], 而基于大样本数据的智能体策略学习在泛化性上表现出极大的优势[94]. 相关研究包括知识牵引与数据驱动[95]、先验知识启发的多智能体双层强化学习方法[96], 面向兵棋推演的多智能体策略协同研究框架[97] 等. 随着大规模群体智能的演化, 融合学习机制的 MARL 有效解决了传统算法在大规模问题中的信度分配难、可拓展性差固有难题, 然而相关研究在现实中的应用还不够成熟, 因此, 如何构建基于学习机制的 MARL, 并应用到海量智能体以及更复杂的环境中已成为未来研究的重要方向.

### 4.3 可解释的多智能体学习方法

现有的基于学习机制的多智能体强化学习算法虽然在很多领域都取得了突破, 但由于算法缺乏可解释性, 限制了模型在安全敏感等领域的应用. 近年来, 研究人员对可解释的多智能体强化学习进行了大量探索, 如用沙普利值解释多智能体强化学习中的合作策略和智能体贡献[98], 利用隐空间多层图(Multiplex latent graphs)来建模智能体之间的交互关系[99] 等, 但总体而言, 这一研究方向目前才刚刚兴起, 在理论上还有诸多不完善之处. 博弈论、MARL 与学习机制的交叉研究在一些新兴领域中得到了可观的应用, 博弈论自诞生以来, 已在经济学、社会、物理学、认知科学和计算机科学等研究领域中取得了大量卓有成效的成果. 近年来,伴随着 AI 技术的快速发展,人们也见证着一些新兴交叉研究领域的兴起,比如社会智能[100−101]、机器智能[102−103]、合作智能[104]、AI 安全[105] 和 AI 伦理[106] 等. 考虑到这些研究主题本身具有广泛的学科交叉性, 如何将博弈论、MARL 与学习机制的交叉研究成果应用于这些新兴领域是未来一个十分值得探究的课题.

## 5 结语

本文从学习机制的角度出发, 梳理了近年来出现的学习机制和多智能体强化学习的融合算法. 为了解决多智能体强化学习中维度爆炸、难以迁移等固有挑战, 很多学者尝试将一些学习机制应用到 MARL 中以改善算法性能. 本文重点介绍了课程学习、演化学习、元学习、分层学习和迁移学习等学习机制与强化学习融合而成的算法, 与经典的多智能体强化学习算法相比, 融合算法在计算效率、少样本训练效果和策略生成等方面有了明显的提升. 在未来多智能体博弈决策当中, 如何“学会学习”在更加科学、合理的多智能体决策中发挥着越来越关键的作用. 融合学习机制的多智能体强化学习通过模仿人类的学习过程, 引导多智能体博弈决策系统涌现出群体智能. 总之, 基于学习机制的多智能体强化学习算法是一个非常有前景的研究方向, 随着这一领域理论和方法的不断发展, 可以预见, “解决智能, 并用智能解决一切”这一目标终将实现.

## 参 考 文 献


[1] Hu X F. *Science of War: The Scientific Basis and Thinking Method of Knowing and Understanding War*. Beijing: Science Press, 2018
(胡晓峰. 战争科学论: 认识和理解战争的科学基础与思维方法. 北京: 科学出版社, 2018)

[2] Sutton R S, Barto A G. *Reinforcement Learning: An introduction*. Cambridge: MIT Press, 2018

[3] Gronauer S, Diepold K. Multi-agent deep reinforcement learning: A survey. *Artif Intell Rev*, 2022, 55(2): 895

[4] Wang Y T, Xue K, Qian C. Evolutionary diversity optimization with clustering-based selection for reinforcement learning//*International Conference on Learning Representations*. Vienna, 2021

[5] Yang Y, Wang J. An overview of multi-agent reinforcement learning from game theoretical perspective [J/OL]. *arXiv preprint* (2021-03-18) [2023-08-08]. https://arxiv.org/abs/2011.00583

[6] Zhang K Q, Yang Z R, Başar T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, 2021: 321

[7] Oroojlooy A, Hajinezhad D. A review of cooperative multi-agent deep reinforcement learning. *Appl Intell*, 2023, 53(11): 13677

[8] Hernandez-Leal P, Kaisers M, Baarslag T, et al. A survey of learning in multiagent environments: Dealing with non-stationarity [J/OL]. *arXiv preprint* (2019-03-11) [2023-08-08]. https://arxiv.org/abs/1707.09183

[9] Da Silva F L, Costa A H R. A survey on transfer learning for multiagent reinforcement learning systems. *Jair*, 2019, 64: 645

[10] Bloembergen D, Tuyls K, Hennes D, et al. Evolutionary dynamics of multi-agent learning: A survey. *Jair*, 2015, 53: 659

[11] Silver D, Hubert T, Schrittwieser J, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 2018, 362(6419): 1140

[12] Jaderberg M, Czarnecki W M, Dunning I, et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, 2019, 364(6443): 859

[13] Hessel M, Modayil J, Van Hasselt H, et al. Rainbow: Combining

improvements in deep reinforcement learning//*Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, 32(1): 3215

[14] Lowe R, Wu Y I, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments. *Adv Neural Inf Proc Syst*, 2017, 30

[15] Foerster J, Farquhar G, Afouras T, et al. Counterfactual multi-agent policy gradients//*Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, 32(1): 2974

[16] Bengio Y, Louradour J, Collobert R, et al. Curriculum learning//*Proceedings of the 26th Annual International Conference on Machine Learning*. Long Beach, 2009: 41

[17] Zhou M, Wan Z, Wang H, et al. MALib: A parallel framework for population-based multi-agent reinforcement learning. *J Mach Learn Res*, 2023, 24(150): 1

[18] Zhao W S, Pajarinen J. Self-paced multi-agent reinforcement learning [J/OL]. *arXiv preprint* (2022-05-20) [2023-08-08]. http://arxiv.org/abs/2205.10016

[19] Long Q, Zhou Z H, Gupta A, et al. Evolutionary population curriculum for scaling multi-agent reinforcement learning [J/OL]. *arXiv preprint* (2020-03-23) [2023-08-08]. http://arxiv.org/abs/2003.10423

[20] Yang J C, Nakhaei A, Isele D, et al. CM3: Cooperative multi-goal multi-stage multi-agent reinforcement learning [J/OL]. *arXiv preprint* (2020-01-24) [2023-08-08]. http://arxiv.org/abs/1809.05188

[21] Wang W X, Yang T P, Liu Y, et al. From few to more: Large-scale dynamic multiagent curriculum learning//*Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(5): 7293

[22] Chen J Y, Zhang Y X, Xu Y F, et al. Variational automatic curriculum learning for sparse-reward cooperative multi-agent problems. *Adv Neural Inf Proc Syst*, 2021, 34: 9681

[23] Wang R D, Zheng L T, Qiu W, et al. Towards skilled population curriculum for multi-agent reinforcement learning [J/OL]. *arXiv preprint* (2023-01-07) [2023-08-08]. http://arxiv.org/abs/2302.03429

[24] Narvekar S, Peng B, Leonetti M, et al. Curriculum learning for reinforcement learning domains: A framework and survey. *J Machine Learn Res*, 2020, 21(1): 7382

[25] Du Y Q, Abbeel P, Grover A. It takes four to tango: Multiagent selfplay for automatic curriculum generation [J/OL]. *arXiv preprint* (2022-01-22) [2023-08-08]. http://arxiv.org/abs/2202.10608

[26] Baker B, Kanitscheider I, Markov T, et al. Emergent tool use from multi-agent autocurricula [J/OL]. *arXiv preprint* (2020-02-11) [2023-08-08]. http://arxiv.org/abs/1909.07528

[27] Yang Y D, Luo J, Wen Y, et al. Diverse auto-curriculum is critical for successful real-world multiagent learning systems [J/OL]. *arXiv preprint* (2020-02-16) [2023-08-08]. http://arxiv.org/abs/2102.07659

[28] Smith J M, Price G R. The logic of animal conflict. *Nature*, 1973, 246: 15

[29] Hilbe C, Šimsa Š, Chatterjee K, et al. Evolution of cooperation in stochastic games. *Nature*, 2018, 559: 246

[30] Such F P, Madhavan V, Conti E, et al. Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning [J/OL]. *arXiv preprint* (2020-04-20) [2023-08-08]. http://arxiv.org/abs/1712.06567

[31] Moriarty D E, Schultz A C, Grefenstette J J. Evolutionary algorithms for reinforcement learning. *Jair*, 1999, 11: 241

[32] Khadka S, Majumdar S, Tumer K. Evolutionary reinforcement learning for sample-efficient multiagent coordination [J/OL]. *arXiv preprint* (2020-06-11) [2023-08-08]. http://arxiv.org/abs/1906.07315

[33] Zhang X W, Clune J, Stanley K O. On the relationship between the OpenAI evolution strategy and stochastic gradient descent [J/OL]. *arXiv preprint* (2017-12-18) [2023-08-08]. http://arxiv.org/abs/1712.06564

[34] Gomes J, Mariano P, Christensen A L. Dynamic team heterogeneity in cooperative coevolutionary algorithms. *IEEE Trans Evol Comput*, 2018, 22(6): 934

[35] Jaderberg M, Dalibard V, Osindero S, et al. Population based training of neural networks[J/OL]. arXiv preprint (2017-01-27) [2017-02-28]. https://arxiv.org/abs/1711.09846

[36] Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 2019, 575(7782): 350

[37] Conti E, Madhavan V, Such F P, et al. Improving exploration in evolution strategies for deep reinforcement learning via a population of novelty-seeking agents. *Adv Neural Inf Proc Syst*, 2018, 31

[38] Liu Z X, Chen B M, Zhou H Y, et al. MAPPER: Multi-agent path planning with evolutionary reinforcement learning in mixed dynamic environments//2020 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*). Las Vegas, 2020: 11748

[39] Fakoor R, Chaudhari P, Soatto S, et al. Meta-Q-learning [J/OL]. arXiv preprint (2020-04-04) [2023-08-08]. http://arxiv.org/abs/1910.00125

[40] Kirsch L, Harrison J, Sohl-Dickstein J, et al. General-purpose In-context learning by meta-learning transformers [J/OL]. *arXiv preprint* (2022-12-08) [2023-08-08]. http://arxiv.org/abs/2212.04458

[41] Doshi-Velez F, Konidaris G. Hidden parameter Markov decision processes: A semiparametric regression approach for discovering latent task parametrizations [J/OL]. *arXiv preprint* (2013-08-12) [2023-08-08]. http://arxiv.org/abs/1308.3513

[42] Wang J X, Kurth-Nelson Z, Tirumala D, et al. Learning to reinforcement learn [J/OL]. *arXiv preprint* (2017-01-23) [2023-

08-08]. http://arxiv.org/abs/1611.05763

[43] Dorfman R, Shenfeld I, Tamar A. Offline meta reinforcement learning–identifiability challenges and effective data collection strategies. *Adv Neural Inf Proc Syst*. 2021, 34: 4607

[44] Sodhani S, Zhang A, Pineau J. Multi-task reinforcement learning with context-based representations//*International Conference on Machine Learning*. Online, 2021: 9767

[45] Fu H T, Tang H Y, Hao J Y, et al. Towards effective context for meta-reinforcement learning: An approach based on contrastive learning//*Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35(8): 7457

[46] Da Silva B C, Basso E W, Bazzan A L C, et al. Dealing with non-stationary environments using context detection//*Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, 2006: 217

[47] Vezhnevets A S, Wu Y H, Leblond R, et al. Options as responses: Grounding behavioural hierarchies in multi-agent RL [J/OL]. *arXiv preprint* (2017-07-10) [2023-08-08]. http://arxiv.org/abs/1906.01470

[48] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks//*Proceedings of the 34th International Conference on Machine Learning*. Sydney, 2017: 1126

[49] Kim D K, Liu M, Riemer M D, et al. A policy gradient algorithm for learning to learn in multiagent reinforcement learning//*International Conference on Machine Learning*. Online, 2021: 5541

[50] Wu Z, Li K, Xu H, et al. L2E: Learning to exploit your opponent//2022 *International Joint Conference on Neural Networks* (*IJCNN*). Padua, 2022: 1

[51] Al-Shedivat M, Bansal T, Burda Y, et al. Continuous adaptation via meta-learning in nonstationary and competitive environments [J/OL]. *arXiv preprint* (2018-02-23) [2023-08-08]. http://arxiv.org/abs/1710.03641

[52] Foerster J N, Chen R Y, Al-Shedivat M, et al. Learning with opponent-learning awareness [J/OL]. *arXiv preprint* (2018-09-19) [2023-08-08]. http://arxiv.org/abs/1709.04326

[53] Sutton R S, Precup D, Singh S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artif Intell*, 1999, 112(1-2): 181

[54] Tang Z T, Zhao D B, Zhu Y H, et al. Reinforcement learning for build-order production in StarCraft II//2018 *Eighth International Conference on Information Science and Technology* (*ICIST*). Cordoba, 2018: 153

[55] Zhang Z J, Li H Z, Zhang L, et al. Hierarchical reinforcement learning for multi-agent MOBA game [J/OL]. *arXiv preprint* (2019-06-21) [2023-08-08]. http://arxiv.org/abs/1901.08004

[56] Liang Z X, Cao J N, Jiang S, et al. Hierarchical reinforcement learning with opponent modeling for distributed multi-agent cooperation//2022 *IEEE 42nd International Conference on Distributed Computing Systems* (*ICDCS*). Bologna, 2022: 884

[57] Bacon P L, Harb J, Precup D. The option-critic architecture//*Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. *San Francisco*, 2017, 31(1): 1726

[58] Precup D, Sutton R S. Multi-time models for temporally abstract planning//*Proceedings of the 10th International Conference on Neural Information Processing Systems*. Denver, 1997: 1050

[59] Liu C Y, Tan Y Q, Liu C A, et al. Application of multi-agent reinforcement learning in robot soccer. *Acta Electron Sin*, 2010, 38(8): 1958
(刘春阳, 谭应清, 柳长安, 等. 多智能体强化学习在足球机器人中的研究与应用. 电子学报, 2010, 38(8): 1958)

[60] Tian Y D, Gong Q C, Shang W L, et al. ELF: An extensive, lightweight and flexible research platform for real-time strategy games//*Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, 2017: 2656

[61] Nachum O, Gu S X, Lee H, et al. Data-efficient hierarchical reinforcement learning//*Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Montréal, 2018: 3307

[62] Levy A, Konidaris G, Platt R, et al. Learning multi-level hierarchies with hindsight [J/OL]. *arXiv preprint* (2019-09-03) [2023-08-08]. http://arxiv.org/abs/1712.00948

[63] Menache I, Mannor S, Shimkin N. Q-cut—Dynamic discovery of sub-goals in reinforcement learning//*Proceedings of the 13th European Conference on Machine Learning*. Berlin, 2002: 295

[64] Kulkarni T D, Narasimhan K, Saeedi A, et al. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. *Adv Neural Inf Proc Syst*, 2016, 29

[65] Xu J, Zhong F W, Wang Y Z. Learning multi-agent coordination for enhancing target coverage in directional sensor networks. *Adv Neural Inf Proc Syst*, 2020, 33: 10053

[66] Ma J, Wu F. Feudal multi-agent deep reinforcement learning for traffic signal control//*Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems* (*AAMAS*). 2020: 816

[67] Vezhnevets A S, Osindero S, Schaul T, et al. FeUdal networks for hierarchical reinforcement learning//*Proceedings of the 34th International Conference on Machine Learning*. Sydney, 2017: 3540

[68] Weiss K, Khoshgoftaar T M, Wang D D. A survey of transfer learning. *J Big Data*, 2016, 3(1): 9

[69] Geng Y Z, Liu E W, Wang R, et al. Hierarchical reinforcement learning for relay selection and power optimization in two-hop cooperative relay network. *IEEE Trans Commun*, 2022, 70(1): 171

[70] Ren T, Niu J W, Dai B, et al. Enabling efficient scheduling in large-scale UAV-assisted mobile-edge computing via hierarchical

reinforcement learning. *IEEE Internet Things J*, 2022, 9(10): 7095

[71] Pan S J, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*, 2010, 22(10): 1345

[72] Zhu Z D, Lin K X, Jain A K, et al. Transfer learning in deep reinforcement learning: A survey [J/OL]. *arXiv preprint* (2023-07-04) [2023-08-08]. http://arxiv.org/abs/2009.07888

[73] Lazaric A. *Transfer in Reinforcement Learning: A Framework and A Survey*. Berlin: Springer, 2012

[74] Taylor A, Dusparic I, Guériau M, et al. Parallel transfer learning in multi-agent systems: What, when and how to transfer?//2019 *International Joint Conference on Neural Networks* (*IJCNN*). Budapest, 2019: 1

[75] MacGlashan J, Ho M K, Loftin R, et al. Interactive learning from policy-dependent human feedback//*Proceedings of the* 34*th International Conference on Machine Learning*. Sydney, 2017: 2285

[76] Li B, Bai S X, Liang S Y, et al. Manoeuvre decision—Making of unmanned aerial vehicles in air combat based on an expert actor-based soft actor critic algorithm. *CAAI Trans Intell Technol*, 2023, 8(4): 1608

[77] Abramson J, Ahuja A, Carnevale F, et al. Improving multimodal interactive agents with reinforcement learning from human feedback [J/OL]. *arXiv preprint* (2022-11-21) [2023-08-08]. http://arxiv.org/abs/2211.11602

[78] Liu W Z, Dong L, Liu J, et al. Knowledge transfer in multi-agent reinforcement learning with incremental number of agents. *J Syst Eng Electron*, 2022, 33(2): 447

[79] Hu F H, Deng Y S, Hamid Aghvami A. Scalable multi-agent reinforcement learning for dynamic coordinated multipoint clustering. *IEEE Trans Commun*, 2023, 71(1): 101

[80] Liang X, Liu Y, Chen T, et al. *Federated Transfer Reinforcement Learning for Autonomous Driving*. Berlin: Springer International Publishing, 2022

[81] Li J C, Shi H B, Hwang K S. Using fuzzy logic to learn abstract policies in large-scale multiagent reinforcement learning. *IEEE Trans Fuzzy Syst*, 2022, 30(12): 5211

[82] Kono H, Kamimura A, Tomita K, et al. Transfer learning method using ontology for heterogeneous multi-agent reinforcement learning. *Int J Adv Comput Sci Appl*, 2014, 5: 10

[83] Da Silva F L, Warnell G, Costa A H R, et al. Agents teaching agents: A survey on inter-agent transfer learning. *Auton Agents Multi Agent Syst*, 2019, 34(1): 9

[84] Zhou L W, Yang P, Chen C L, et al. Multiagent reinforcement learning with sparse interactions by negotiation and knowledge transfer. *IEEE Trans Cybern*, 2017, 47(5): 1238

[85] Li D L, Wang J P. FedMD: Heterogenous federated learning via model distillation [J/OL]. *arXiv preprint* (2019-10-08) [2023-08-08]. http://arxiv.org/abs/1910.03581

[86] Shi H B, Li J C, Mao J H, et al. Lateral transfer learning for multiagent reinforcement learning. *IEEE Trans Cybern*, 2023, 53(3): 1699

[87] Lin K X, Zhao R Y, Xu Z, et al. Efficient large-scale fleet management *via* multi-agent deep reinforcement learning// *Proceedings of the* 24*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. London, 2018: 1774

[88] Liu X Y, Tan Y. Attentive relational state representation in decentralized multiagent reinforcement learning. *IEEE Trans Cybern*, 2022, 52(1): 252

[89] Schmid K, Belzner L, Gabor T, et al. Action markets in deep multi-agent reinforcement learning//*International Conference on Artificial Neural Networks*. Rhodes, 2018: 240

[90] Ghosh D, Rahme J, Kumar A, et al. Why generalization in RL is difficult: Epistemic POMDPs and implicit partial observability. *Adv Neural Inf Proc Syst*, 2021, 34: 35502

[91] Yang Y D, Luo R, Li M, et al. Mean field multi-agent reinforcement learning//*International Conference on Machine Learning*. Jinan, 2018: 5571

[92] Huang W L, Mordatch I, Pathak D. One policy to control them all: Shared modular policies for agent-agnostic control// *Proceedings of the* 37*th International Conference on Machine Learning*. Vienna, 2020: 4455

[93] Muller P, Omidshafiei S, Rowland M, et al. A generalized training approach for multiagent learning//8*th International Conference on Learning Representations*. Ababa, 2020: 1

[94] Team O E L, Stooke A, Mahajan A, et al. Open-ended learning leads to generally capable agents [J/OL]. *arXiv preprint* (2021-07-31) [2023-08-08]. http://arxiv.org/abs/2107.12808

[95] Cheng K, Chen G, Yu X H, et al. Knowledge traction and data-driven wargame AI design and key technologies. *Syst Eng Electron*, 2021, 43(10): 2911
(程恺, 陈刚, 余晓晗, 等. 知识牵引与数据驱动的兵棋 AI 设计及关键技术. 系统工程与电子技术, 2021, 43(10): 2911)

[96] Chen X X, Huang K H, Liang X X, et al. Tactical prior knowledge inspiring multi-agent bilevel reinforcement learning. *J Command Contr*, 2022, 8(1): 72
(陈晓轩, 黄魁华, 梁星星, 等. 战术先验知识启发的多智能体双层强化学习. 指挥与控制学报, 2022, 8(1): 72)

[97] Weng J Y, Chen H Y, Yan D, et al. Tianshou: A highly modularized deep reinforcement learning library. *J Mach Learn Res*, 2022, 23(1): 12275

[98] Heuillet A, Couthouis F, Díaz-Rodríguez N. Collective eXplainable AI: Explaining cooperative strategies and agent contribution in multiagent reinforcement learning with shapley values. *IEEE Comput Intell Mag*, 2022, 17(1): 59

[99] Sun F Y, Kauvar I, Zhang R, et al. Interaction modeling with multiplex attention. *Adv Neural Inf Proc Syst*, 2022, 35: 20038

[100] Vouros G A. Explainable deep reinforcement learning: State of

the art and challenges. *ACM Comput Surv*, 2022, 55(5): 92

[101] Liu X, Liu S Y, Zhuang Y K, et al. Explainable reinforcement learning: Basic problems exploration and method survey. *J Softw*, 2023, 34(5): 2300
(刘潇, 刘书洋, 庄韫恺, 等. 强化学习可解释性基础问题探索和方法综述. 软件学报, 2023, 34(5): 2300)

[102] Qing Y P, Liu S Y, Song J, et al. A survey on explainable reinforcement learning: Concepts, algorithms, challenges [J/OL]. arXiv preprint (2022-11-12) [2023-08-08]. http://arxiv.org/abs/2211.06665

[103] Papoudakis G, Christianos F, Rahman A, et al. Dealing with non-stationarity in multi-agent deep reinforcement learning [J/OL]. *arXiv preprint* (2019-06-11) [2023-08-08]. http://arxiv.org/abs/1906.04737

[104] Foerster J N, Assael Y M, de Freitas N, et al. Learning to communicate with deep multi-agent reinforcement learning. *Adv Neural Inf Proc Syst*, 2016: 2137

[105] Cichocki A, Kuleshov A P. Future trends for human-AI collaboration: A comprehensive taxonomy of AI/AGI using multiple intelligences and learning styles. *Comput Intell Neurosci*, 2021, 2021: 8893795

[106] Mirowski P, Pascanu R, Viola F, et al. Learning to navigate in complex environments [J/OL]. *arXiv preprint* (2017-01-17) [2023-08-08]. http://arxiv.org/abs/1611.03673