

群视角下的多智能体强化学习方法综述

项凤涛, 罗俊仁, 谷学强, 苏炯铭, 张万鹏

(国防科技大学智能科学学院, 湖南 长沙 410073)

摘要: 多智能体系统是分布式人工智能领域的前沿研究概念, 传统的多智能体强化学习方法主要聚焦群体行为涌现、多智能体合作与协调、智能体间交流与通信、对手建模与预测等主题, 但依然面临环境部分可观、对手策略非平稳、决策空间维度高、信用分配难理解等难题, 如何设计满足智能体数量规模比较大、适应多类不同应用场景的多智能体强化学习方法是该领域的前沿课题。首先简述了多智能体强化学习的相关研究进展; 其次着重从规模可扩展与种群自适应两个视角对多种类、多范式的多智能体学习方法进行了综合概述归纳, 系统梳理了集合置换不变性、注意力机制、图与网络理论、平均场理论共四大类规模可扩展学习方法, 迁移学习、课程学习、元学习、元博弈共四大类种群自适应强化学习方法, 给出典型应用场景; 最后从基准平台开发、双层优化架构、对抗策略学习、人机协同价值对齐和自适应博弈决策环共5个方面进行了前沿研究方向展望, 该研究可为多模态环境下多智能体强化学习的相关前沿重点问题研究提供参考。

关键词: 分布式智能; 平均场理论; 图神经网络; 元学习; 元博弈

中图分类号: TP39

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202326

Survey on multi-agent reinforcement learning methods from the perspective of population

XIANG Fengtao, LUO Junren, GU Xueqiang, SU Jiongming, ZHANG Wanpeng

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China

Abstract: Multi-agent systems are a cutting-edge research concept in the field of distributed artificial intelligence. Traditional multi-agent reinforcement learning methods mainly focus on topics such as group behavior emergence, multi-agent cooperation and coordination, communication and communication between agents, opponent modeling and prediction. However, they still face challenges such as observable environment, non-stationary opponent strategies, high dimensionality of decision space, and difficulty in understanding credit allocation. How to design multi-agent reinforcement learning methods that meet the large number and scale of intelligent agents and adapt to multiple different application scenarios is a cutting-edge topic in this field. This article first outlined the relevant research progress of multi-agent reinforcement learning. Secondly, a comprehensive overview and induction of multi-agent learning methods with multiple types and paradigms were conducted from the perspectives of scalability and population adaptation. Four major categories of scalable learning methods were systematically sorted out, including set permutation invariance, attention, graph and network theory, and mean field theory. There were four major categories of population adaptive reinforcement learning methods: transfer learning, course learning, meta learning, and meta game, and typical application scenarios were provided. Finally, the frontier research directions were prospected from five aspects: benchmark platform development, two-layer optimiza-

收稿日期: 2023-07-21; 修回日期: 2023-08-22

通信作者: 项凤涛, xiangfengtao@nudt.edu.cn

tion architecture, adversarial strategy learning, human-machine collaborative value alignment and adaptive game decision-making loop, providing reference for the research on relevant frontier key issues of multi-agent reinforcement learning in multimodal environments.

Key words: distributed intelligence, mean field theory, graph neural network, meta learning, meta game

0 引言

多智能体系统 (multi-agent system, MAS) 是指在同一环境中由多个交互智能体组成的系统, 常用来解决独立智能体或单层系统难以解决的问题, 其中的智能可通过知识推理、交互学习等方式获得。多智能体深度强化学习方法结合了深度学习、强化学习与多智能体系统理论等技术, 赋予智能体更强的感知、推理、决策和学习能力, 并在许多应用场景中展现出强大的潜力。目前, 国内外相关领域学者从不同视角对多智能体强化学习的相关研究进行了综述分析。Yang 等人^[1]借助博弈理论相关知识, 研究了基于博弈 (特别是元博弈) 理论的多智能体学习方法; Silva 等人^[2]从策略迁移的角度对多智能体自身内部和多智能体间的策略迁移问题进行了综述分析; Yang 等人^[3]从认知和偶然不确定性, 内生新颖性和影响性等方面分析了多智能体策略探索问题; 王涵等人^[4]从值函数、直接策略搜索、通信效率提升和应急通信4个方面梳理总结了基于通信的多智能体强化学习方法; 殷昌盛等人^[5]从分层学习的角度对 Option、HAM、Max-Q 等多智能体强化学习方法进行了梳理分析; 王龙等人^[6]利用跨学科交叉视角综合分析了面向4种博弈形式的多智能体学习方法; 罗俊仁等人^[7]从多智能体博弈模型出发, 区分离线与在线两阶段, 梳理多智能体博弈学习方法。此外, 国内邓小铁课题组为多人一般和随机博弈引入近似马尔可夫完美纳什均衡解概念, 证明了该解的计算复杂度为 PPAD-Complete^[8], 谢广明课题组研究了基于强化学习的多智能体系统控制问题^[9]。

近年来, 随着大数据、大算力、大模型等概念的提出, 大规模系统 (车辆、电力、无人机集群等) 已然成为当前学习类方法的主要研究对象。一些研究从多智能体概念出发提出了许多 (many) 智能体^[10]、大规模 (large scale) 智能体^[11]、大量地 (massively) 多智能体^[12]、可扩展 (scaling、scalable) 多智能体^[13-14]等概念。借助各类计算机博弈及视频游戏平台、数字孪生系统、兵棋推演及模拟

仿真软件, 研究适用于一群智能体的规模可扩展性强化学习方法, 应对多类场景的种群自适应强化学习方法仍然充满挑战。

本文的整体脉络结构如图1所示。首先围绕多智能体强化学习及其面临的挑战, 概述了多智能体强化学习基础和面临的挑战, 简要介绍了多智能体强化学习范式、分布式、协同对抗等前沿研究重点; 其次从“群”视角出发, 区分规模可扩展与种群自适应, 分类归纳多种学习方法, 其中将规模可扩展多智能体强化学习方法分为集合置换不变性、注意力机制、图与网络理论、平均场理论四大类, 将面向种群的自适应强化学习方法分为迁移学习、课程学习、元学习、元博弈四大类; 最后总结了全文, 展望了多智能体深度强化学习方法的应用前景, 讨论了未来发展方向。

1 多智能体强化学习概述

1.1 多智能体强化学习基础

多智能体强化学习通常可直接采用马尔可夫博弈 (Markov game) 模型来建模, 如图2所示, 可表示成八元组 $\langle N, S, A, T, R, O, Z, \gamma \rangle$, 其中: N 表示智能体的数量, S 是所有状态的集合, $S_t \in S$ 表示博弈在时刻 t 的状态, $A = A_1 \times A_2 \times \cdots \times A_N$ 是所有智能体的联合动作的集合, $a \in A$ 是某个特定的联合动作, $a_i^t \in A_i$ 表示第 i 个智能体在 t 时刻采取的动作, $T: S \times A \times S \rightarrow [0, 1]$ 是状态转移概率函数, $R = [R_1, R_2, \cdots, R_N]: S \times A \times S \rightarrow R^N$ 是联合奖励函数, $O = O_1 \times O_2 \times \cdots \times O_N$ 是所有智能体联合观测的集合, $Z(s): S \rightarrow O$ 是观测函数, 控制所有智能体在状态 s^t 时能够感知到的具体观测值, γ 是折扣因子。

相比用博弈论来描述多智能体系统的状态变化, 对于协作型多智能体系统, 由于智能体之间是协作关系, 通常不考虑非协作方智能体行为, 故通常采用分散式马尔可夫决策过程 (Dec-POMDP) 来建模。近年来, 由于深度强化学习技术的日渐成熟, 一些新的研究尝试以深度强化学习为基础, 在多智能体系统中作深度强化学习扩展, 提出了一系

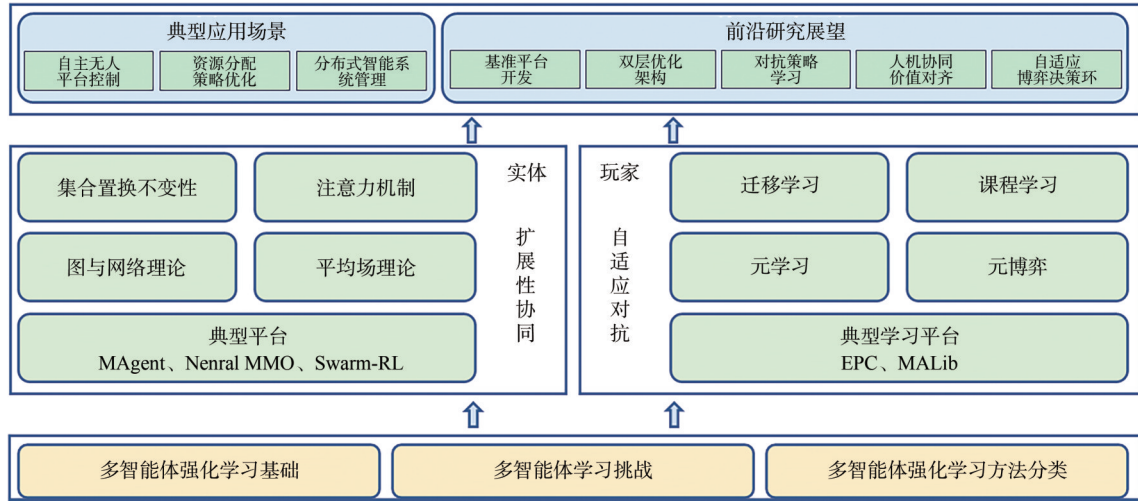


图1 论文整体架构

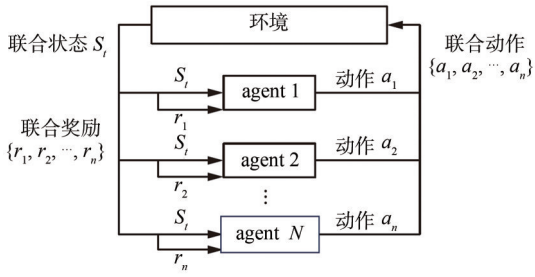


图2 多智能体强化学习示意

列多智能体深度强化学习方法。

自2016年开始，受深度强化学习的成功实践影响，多智能体深度强化学习也得到了快速发展。其中集中式强化学习方法将多智能体系统看成一个具有中央控制单元的单智能体系统，依靠其统一学习和调度，但面临着效率低、维数高等问题。分布式强化学习方法中每个智能体拥有决策能力，每个智能体可以根据自身观察到的环境状态自主决策。根据智能体是否与其他智能体交互，将分布式多智能体强化学习方法分为独立式强化学习方法和协同式强化学习方法，其中协同式强化学习方法在智能体间加入了协调合作机制，让多智能体能够在决策前考虑智能体之间的相互影响，智能体的策略最终满足合理性和收敛性，是目前主流的多智能体深度强化学习方法。

1.1.1 多智能体涌现行为分析

该类问题的相关研究是将单智能体的深度强化学习算法应用在多智能体系统中，分析是否能在动态环境中涌现出新的行为。Zheng等人^[10]设计了面向满足超多智能体群体智能分析的MAgent平台。

Tampuu等人^[15]在乒乓球竞争环境中，使用独立的DQN算法，查看两个智能体能否学会比赛动作。事实证明，原来独立的算法也可以在零和游戏的环境中学习到更好的策略。在独立的多智能体强化学习中，学习自己的策略的同时，固定其他策略，有助于帮助算法收敛，自对弈是一种常见的方法。Leibo等人^[16]提出的Malthusian强化学习方法，使用了自对弈技巧，极大地缓解了环境的不稳定性问题。

1.1.2 多智能体合作与协调

该类问题的研究不考虑或者很少考虑智能体之间是如何进行交流的。相较于学习交流需要智能体之间出现显式的通信信道，此类工作更加注重与多智能体系统自身的协调与协作。通过结合博弈论的相关工作，这部分内容更加关注于多智能体系统本身的数学模型。每一个时间步，合作智能的大部分都是一样的收益。早期的多智能体协作方法大多考虑集中式评估集中执行的思路，但完全集中的学习导致了适应大规模环境的模型过于复杂。Lowe等人^[17]提出了基于DDPG的扩展算法MADDPG，该算法提出了极大降低学习成本的集中训练-分散执行的学习方式。这种集中训练-分散执行的框架也成为范式，大部分的方法都是建立在协作的基础上的。Palmer等人^[18]提出使用“宽容”的想法，设计基于加权的双Q网路来降低过预估，以应对多智能体之间策略的非平稳问题。此外，一些研究关注合作环境中多智能体的信用分配问题，即如何评估每个智能体的动作对当前回报的贡献。基于此，Foerster等人^[19]提出了反事实基线（counterfactual

multi-agent policy gradients, COMA), 评估每个智能体的动作对当前奖励的贡献度, 通过优势函数加速算法收敛。Sunehag 等人^[20]着眼于通过分解值函数来简化中心化的评价网络, 从而提出了值分解网络 (value-decomposition networks, VDN), 将全局的值函数分解为多个智能体本地值函数的累加, 使中心的评价网络进一步简化。随后, Rashid 等人^[21]证明了 QMIX 网络是基于此而提出的, 只要全局值函数关于本地值函数的映射是单调递增的。本地值函数耦合的权重值是用神经网络求得的, 全局值函数是用本地值函数加权求得的, VDN 可以作为一个特殊情况, 用于 QMIX。随着注意力的成熟, 一些研究提出利用注意力来计算全局信息, 同时利用注意力参与值函数的信息融合。基于合作的方法的最大挑战是智能体数量的难扩展性, 一旦环境中智能体数量改变, 算法通常无法适用。

1.1.3 多智能体交流与通信

该类问题的研究是为了帮助智能体之间达成合作交流或竞争的目的, 在自主学习的基础上, 增加可交流模块, 学习何时交流、如何交流等。此部分工作通常考虑部分可观环境中的一组协作智能体, 其中智能体需要通过信息交互来最大化它们的联合回报。Foerster 等人^[22]提出了 RIAL 和 DIAL 算法, 两种方法都是利用神经网络拟合值函数以及智能体之间的交流信息, 同时也在多智能体强化学习中引入了参数共享的概念, 简化了多智能体强化学习的模型复杂度。文献[23]提出了一种名为 CommNet 的网络结构, 通过求和运算将智能体之间的沟通信息聚合起来, 这种方法与图网络中的聚合思路类似, 能够对环境中动态变化的智能体数量进行有效的处理。Peng 等人^[24]提出了一种思想类似于双向循环神经网络的多智能体强化学习方法 BiCNet, 让每个智能体在处理动态数量环境的同时, 拥有全局观察的能力。此外, 一些研究使用注意力机制的通信结构, 能够在智能体间智能地选择通信关系, 并对通信信息进行计算。

1.1.4 多智能体建模与预测

该类问题的研究主要聚焦于对对手的策略建模, 其关注点在于对多智能体系统中的其他智能体的策略进行预测。通过这种显式的策略评估和建模, 让智能体之间更协调地合作是这部分工作的最终目的。在该研究方向中, 智能体对其他智能体的策略进行建模, 并对其他智能体的行为进行推

断^[25]。深度强化对手网络 (deep reinforce opponent network, DRON)^[25]是最早提出利用深度神经网络建模智能体的研究工作; MADDPG 算法^[17]基于多智能体 Actor-Critic 框架, 利用集中式 critic 网络对自身行为进行评估, 结合其他智能体的信息来帮助自己的 actor 网络学习策略。对手学习意识的学习 (learning with opponent-learning awareness, LOLA)^[26]则通过引入新的学习规则预测对手策略参数的更新, 更好地对应预测行为。对手建模技术的引入解决了对抗环境下对手策略参数未知的问题, 通过对对手状态-行动轨迹的观察, 采用最大似然估计值来求得对手策略参数的估计值, 但对于风格复杂多变的对手往往会难以应付。Rabinowitz 等人^[27]提出了心智理论神经网络 (theory of mind network, ToM-net), 使机器能够学习他人的心理状态, 通过观察智慧体的行为, 利用元学习法对其进行建模, 从而获得一个具有强大的智慧体行为先验知识的模型。该模型可以更丰富地预测智能体特征和心理状态, 使用少量的行为观测。Li 等人^[28]从鲁棒强化学习中引入极大极小思想, 对 MADDPG 算法进行了扩展, 提出了极大极小多智能体深度确定策略梯度 (minimax multi-agent deep deterministic policy gradient, M3DDPG) 算法, 假设环境中的其他智能体都会对自身产生负面影响, 利用“最坏噪声”提升智能体的鲁棒性。

1.2 多智能体强化学习挑战

正是由于多个智能体的存在, 多智能体深度强化学习面临着一系列挑战。

(1) 环境部分可观。受限于物理空间上的分离与自身局部感知能力, 单个智能体仅能观测整个环境状态的一部分, 故整体的状态空间是部分可观的。

(2) 策略非平稳。多智能系统中的每个智能体都只能控制自己的策略, 并且各方策略之间没有强制的约束关系, 导致状态转移和奖励分配函数无法收敛到一个平滑的分布上, 从而导致每个智能体的新状态和获得的奖励不再具有确定性。

(3) 维度灾难性。更大的动作空间和规模可扩展的智能体, 使得联合动作空间呈指数级增长, 总体上具有巨大的维度。

(4) 信用分配不合理。对于协同的多智能系统来说, 由于环境反馈的奖励信号是由多个智能体的共同行动决定的, 需要杜绝滥竽充数的“懒”智能

体, 让每个智能体的奖励分配符合真正的贡献度, 所以信用分配的合理性面临挑战。

(5) 环境探索难。探索与利用是强化学习的基本问题, 分别是基于当前环境状态选择新动作来执行和利用已经探索过的动作选择最佳动作。

1.3 多智能体强化学习方法分类

1.3.1 学习范式

随着新算法的支撑及新算力的突破, 多智能体强化学习的范式也正在发生转变。

(1) 独立式学习, 每个智能体采用完全独立的方式进行策略学习, 但由于智能体之间没有任务信息共享, 实际性能往往很不好。

(2) 集中式学习, 多个智能体的策略训练与动作执行全部采用集中式的网络, 而非显式地拆分成属于单个智能体的子网络。由于通常采用信息共享的方式, 故不能很好地扩展到分布式环境。

(3) 分布式学习, 每个智能体仅根据自身观测信息、网络近邻智能体信息或通信消息智能体信息, 就可独立做出决策动作。

1.3.2 协作方式

通常这类方法中策略训练是集中式的, 动作执行是分布式, 而并非所有阶段都是分布式的。围绕策略训练与动作执行, 现有多智能体强化学习可划分为四大类。

(1) 基于集中式评估的学习方法, 采用独立式行动执行和集中式的策略学习评估。由于注意力机制可以很好地区分不同的输入, 可以利用注意力机制从高维输入信息中寻找与智能体决策最相关的信息。相关典型工作有 MADDPG^[29]、COMA^[19]、基于注意力机制的队友联合策略建模方法 ATT-MADDPG^[30]。

(2) 基于联合动作值分解的学习方法, 将多智能体系统的联合动作值函数分解成某一特定组合的个体动作值函数, 按机制的不同可分为简单因子分解型、个体全局最大 (individual global max, IGM) 型和注意力机制型^[31]。

(3) 基于协作交互的学习方法, 将智能体之间的协作交互关系纳入考虑, 划定智能体的领域 (agent neighborhood), 相关典型工作如利用协作图 (coordination graph) 来描述智能体之间的密切关系^[32-33]。

(4) 基于联网通信学习方法, 利用连接多个智能体之间的通信信息进行学习。根据通信机制的不同, 可分为以直接策略搜索为基础的方法、以值函数为基础的方法、以提高通信效率为导向的方法、以应急通信为导向的方法等^[4]。

1.3.3 群视角下的协同对抗

对于大规模多智能体系统, 处理数量和规模方面的动态变化是当前深度强化学习方法面临的突出挑战。基于学习 (深度学习、强化学习) 设计的迭代式问题求解方法是离线策略学习的基础范式。由于环境及对手的非平稳性, 离线训练的蓝图策略通常很难直接运用于在线对抗。在线博弈对抗过程与离线利用模拟多次对抗学习博弈过程不同, 博弈各方处于策略解耦合状态, 与离线批 (batch) 式策略学习方法不同, 在线博弈对抗策略的求解本质是一个流 (flow) 式学习过程, 需要根据少量此前交互样本来做决策。基于分布式深度强化学习, Yin 等人^[34]提出了多玩家多智能体学习工具箱。

多智能体协同对抗通常包括两层, 即内层的多实体协同与顶层的玩家对抗, 面向多智能体协同对抗的双层学习框架如图3所示。

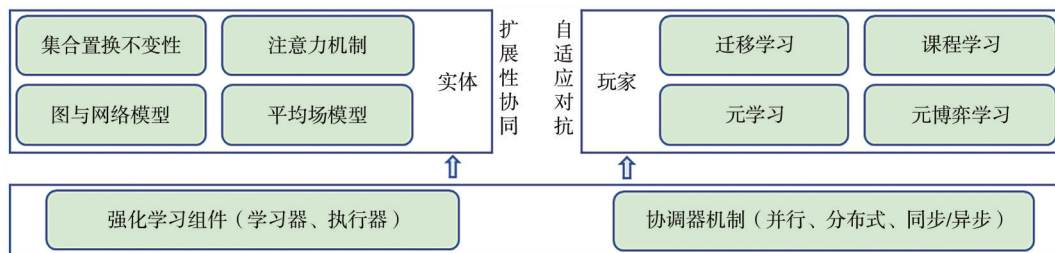


图3 多智能体协同对抗双层学习框架

2 规模可扩展学习方法

传统的多智能体强化学习方法对于拥有少量智

能体的应用场景效果比较明显, 但对于智能体数据规模过大场景, 指数爆炸导致学习时的样本效率过低。如图4所示, MAgent 对抗场景中, 智能体的

数据规模比较大,为了应对大规模智能体,设计规模可扩展的多智能体强化学习方法已然成为当前研究的焦点。

依靠分布式训练框架 IMPALA, DeepMind 在开发星际争霸 AlphaStar 时,采用了集中式训练分

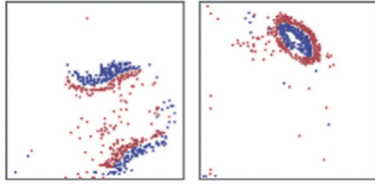


图4 MAgent 对抗场景^[10]

布式执行的范式设计了三大类智能体对象:主智能体(main agent)为正在训练的智能体及历史数据,采用优先级虚拟自对弈的方式来选取;联盟利用者(league exploiter)能打败联盟中的所有智能体,按照有优先级虚拟自对弈的方式与全联盟的对手进行训练;主利用者(main exploiter)能够打败所有的智能体。

规模可扩展学习方法主要聚焦在如何为深度网络的状态输入设计满足动态变化(变长或变维)的状态表示,这是这类问题的本质。当前主流建模框架与典型方法如表1所示。

表1 相关建模框架与典型可扩展方法

类别	建模框架与典型方法	主要机制及方法特点	缺点分析
集合置换不变性	深度集合 ^[35]	利用置换不变性设计满足集合数据处理的网络	忽略了智能体之间的差异,无法满足异构智能体
	深度集合Q值 ^[36]	在深度强化学习中引入深度集合网络架构	
	深度集群 ^[37-38]	多无人机交互采用深度集合架构编码	
	深度集合平均场 ^[39]	基于深度集合架构与平均场理论设计PPO算法	
注意力机制	注意力关系型编码器 ^[40]	可以聚合任意数量邻近智能体的特征表示	忽略了智能体之间的信息交互,无法保证有效区分各智能体的重要性
	随机实体分解 ^[41]	采用随机分解来处理不同类型的和数量的智能体	
	基于注意力的深度集合 ^[42]	采用基于注意力的深度集合框架来控制集群	
	通用策略分解Transformer ^[43]	利用Transformer来学习分组实体的策略	
图与网络理论	种群不变Transformer ^[44]	设计具备种群数量规模不变的Transformer	需要建模复杂的交互关系,很难处理超大规模智能体策略空间的指数增长
	可分解马尔可夫决策 ^[32]	动态贝叶斯网络分解状态变换函数	
	联网分布式决策 ^[45]	利用图表示部分可观的局部交互	
	可分解分散式决策 ^[46]	利用协调图将状态变换函数分解成动态贝叶斯网络	
	图卷积网络 ^[47]	直接利用图卷积网络学习大规模无人机编队控制	
	基于聚合的图神经网络 ^[48]	利用聚合操作来处理变长维度的输入	
	图Q值混合网络 ^[49]	利用图神经网络与注意力机制学习值函数分解	
	图注意神经网络 ^[50]	设计小组与个人之间的注意力神经网络	
	深度循环图网络 ^[51]	结合门控循环单元和图注意力网络模型	
	深度协调图 ^[52]	设计基于超图表征智能体关系的图卷积神经网络	
	超图卷积混合 ^[53]	基于超图卷积的值分解	
平均场理论	协作图贝叶斯博弈 ^[54]	构建满足智能体之间交互的非平稳交互图	忽略了数量规模信息和智能体重要程度
	平均场多智能体强化学习 ^[55]	基于平均嵌入设计满足多智能体的平均场Q学习	
	平均场博弈 ^[56-57]	基于智能体与邻居的交互图构建局部平均场博弈	
	平均场控制 ^[58]	将多智能体强化学习转换成高维单智能体决策	

2.1 基于集合置换不变性的方法

可扩展性本质是要求降低方法对输入的敏感度,与等变性(equivariant)不同,集合具备置换不变性(permutation invariance)的特性。Zaheer等人^[35]提出的深度集合(deep sets)是一类面向集合的深度强化学习框架。深度集合Q值^[36]设计了面向集合的深度强化学习框架。深度集群网络^[37-38]是一类面向无人机集群控制、规划的强化学习框架,实现

了四旋翼无人机集群的策略虚实迁移。

2.2 基于注意力机制的方法

与集合理论中集合元素置换不变性不同的是,基于注意力机制的方法更加强调智能体角色、重要性的不同。

2.2.1 一般注意力

注意力关系型编码器方法^[40]采用社交注意力池化机制来学习每个邻居节点的重要性程序。随机

实体分解方法^[41]采用注意力机制与QMIX混合方法，利用随机分解来处理不同类型的智能体。基于注意力的深度集合方法^[42]将注意力与深度集合混合，采用基于注意力的深度集合框架来控制集群。

2.2.2 Transformer

Transformer作为一种新的自注意力机制，用于构建注意力机制的学习方式成为一种必然。通用策略分解Transformer方法^[43]利用Transformer来学习分组实体的策略，消除了模型固定输入输出的约束，提升了模型的可扩展性。种群不变Transformer方法^[44]采用种群数量规模不变的Transformer结构。

2.3 基于图与网络理论的方法

为解决多智能体强化学习的组合特性问题，获得高效和可扩展性的算法，一个流行的方向是使用图来表示智能体之间的稀疏交互，因为在许多实际问题中，并不是所有的智能体都彼此交互。图形化建模方法是高度通用的，它允许对智能体之间的交互进行建模，这种建模比标准模型要稀疏得多。

2.3.1 动态贝叶斯网络

一种常见的方法是假设问题的图是可分解结构的，考虑状态变换函数、观测与奖励函数可能具有的独立性特点，将联合函数分解成小因子形式，如图5所示。可分解马尔可夫决策（factored MDPs）模型^[32]是一类基于图分解来实现可扩展性的框架之一。将状态转换模型分解成如图5所示的动态贝叶斯网络（dynamic Bayesian network, DBN）。联网分布式决策（networked distributed POMDPs, ND-POMDPs）模型^[45]是一类融合了部分可观与协调图的统一框架。可分解分散式决策（factored dec-POMDPs）模型^[46]可用于表示基于图的多智能体系统，与可分解马尔可夫决策模型类似，可用动态贝叶斯网络表示状态变换函数和观测模型。

2.3.2 图神经网络

直接利用图结构来描述大规模系统是一类面向现实通联场景的可行方案。图卷积网络^[47]是一类

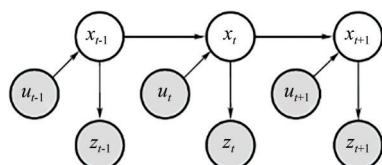


图5 动态贝叶斯网络示意

最早利用图网络设计分布式控制器来实现大规模无人机编队控制的方法。基于聚合的图神经网络^[48]可以处理变长输入，从而适用数量规模比较大的情形。图Q值混合网络^[49]基于QMIX算法，利用图神经网络与注意力机制混合来应对值函数分解与奖励分配。图神经网络^[50]试图通过构建群组内注意力与个体间注意力网络来学习智能体的状态表示。深度循环图网络^[51]采用门控循环单元来处理输入，结合分层图注意力模型，提高模型的可扩展性。

2.3.3 复杂网络模型

复杂网络的相关研究包括小世界网络、渗流原理、动力学模型等。复杂网络理论长期以来专注于利用图理论来理解复杂的大规模系统。近年来，一个重要的新兴领域是研究处理网络节点间的高阶相互作用，已然超越了简单的成对相互作用。深度协调图^[52]使用基于超图的图神经网络来学习智能体之间的交互，由于超图可以表征智能体之间的高阶交互。超图卷积混合网络^[53]采用基于超图卷积的值分解方法。此外，一些研究将重点从图上动态系统转移至动态自适应图系统，同时考虑智能体的变化与智能体交互之间的变化。为了更加真实并适用于现实场景，多智能体系统不仅需要适应不断变化的环境，还需要适应系统中不断变化的交互、操作和连接。协作图贝叶斯博弈（collaborative graphical Bayesian games, CGBG）模型^[54]试图构建满足智能体之间交互的非平稳交互图，智能体在每一步均可更新交互。高阶和自适应网络是表示实际多智能体问题的一种自然而现实的方法。

2.4 基于平均场理论的方法

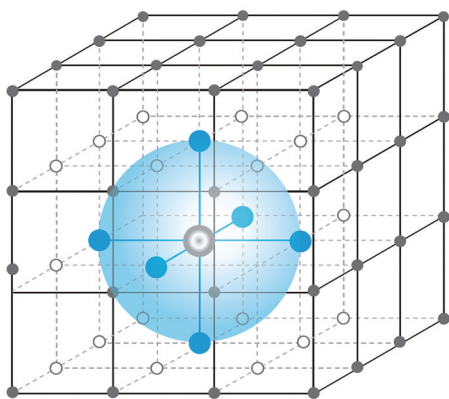
基于平均场理论，利用平均嵌入的思想，计算多个智能体的状态表示平均值，将变规模智能体的状态编码转换成固定维数张量。但此类方法忽略了智能体的数量规模信息和各个智能体的重要程度。

2.4.1 平均场多智能体强化学习

Yang等人^[55]提出了基于平均场理论的多智能体强化学习方法。智能体的Q函数可以通过成对的局部交互分解，如图6所示。每个智能体被表示成格网中的一个节点，接受其邻域智能体的平均作用，多智能体之间的交互转化成两个智能体之间的交互。

2.4.2 平均场博弈

利用平均场理论的思想，多智能体问题被简化为无限智能体极限。直观地说，所有智能体之间的

图6 格网中的智能体节点表示^[55]

相互作用，在平均的博弈中，都被简化为所有智能体的质量、两个机体之间的相互作用，以及任何具有代表性的单个智能体行为。最重要的是，这种简化将一般复杂的多智能体问题简化为竞争场景下的不动点方程或合作场景下的高维单智能体问题。面向平均场博弈的可扩展深度强化学习方法^[56]利用虚拟自对弈与扩展式在线镜像梯度下降来学习平均场博弈均衡。智能体只需与其他智能体的一个子集进行交互，子集可以通过图上邻居来表示。图元（graphon）作为大型图的极限^[57]，常被用来描述图上的平均场博弈，可用于表述连续或离散时间下的静态或动态智能体之间的交互。如图7所示，图元作为邻接矩阵的连续域版本，提供了一种易于处理的建模大型图极限的方法。

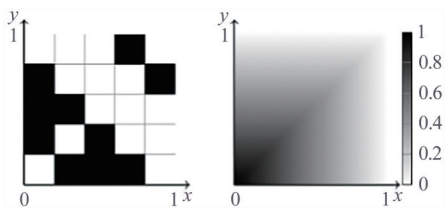


图7 由5个节点构成图（左）的极限图元（右）

2.4.3 平均场控制

与竞争式平均场博弈相比，平均场控制的完全合作框架描述了另一类重要的问题，近年来，协作式平均场控制方法得到了广泛发展。基于Q学习的方法^[58]可有效地学习可扩展的控制策略。平均场博弈与平均场控制的关系如图8所示。

2.5 面向可扩展性的典型学习平台

MAgent^[10]是一个支持多智能体强化学习的研究和开发平台。与以往单一或多智能体强化学习的研究平台不同，MAgent专注于支持需要数百到数

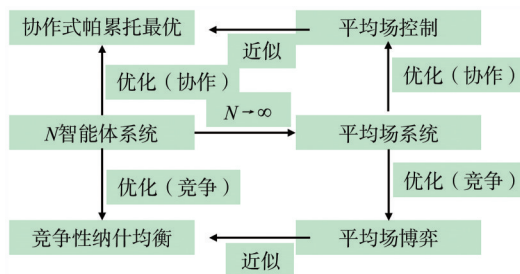


图8 平均场博弈与平均场控制的关系

百万智能体的任务和应用。在一群智能体之间的互动中，它不仅研究智能体最优策略的学习算法，更重要的是，它能够对人工智能社会中出现的个体智能体行为和社会现象进行观察和理解，包括沟通语言、领导能力、利他性等方面。MAgent是高度可扩展的，可以在一个GPU服务器上托管多达100万个智能体。相关示例演示了在MAgent中通过从零开始学习涌现的集体智慧。

Neural MMO^[12]是一个用于人工智能研究的大型多智能体环境。智能体在一个持久的游戏世界中搜寻资源并参与战略战斗。在环境的边缘位置随机生成智能体。它们必须获取食物和水，避免为了维持生命而被别的智能体击倒。踩在森林瓦片上或临水的地方，食物供给或水供给将分别得到重新补充。然而，森林瓦片的食物供应有限，随着时间的推移，食物会缓慢再生。智能体采用近战、远攻、法术3种战斗风格进行战斗。

Swarm-RL^[41]是一类基于端到端深度强化学习分布式控制四旋翼无人机集群的仿真试验平台。学习无人机群控制器的可能性，通过大规模的多智能体到端的强化学习，通过零样本迁移的方式，这些控制器的策略能够迁移到真正的四旋翼无人机上。通过训练神经网络参数化的策略，能够以完全分散的方式控制集群中的单个无人机。仿真实验展示了先进的群集行为，在紧密队形中执行攻击性机动，同时避免相互碰撞，打破和重新建立队形以避免与移动障碍物碰撞，并在“追赶—逃避”任务中有效协作。此外，模拟环境中学习到的模型可以成功部署到真实的四旋翼无人机上。

2.6 可扩展学习方法小结

对于大规模多智能体系统而言，其数量规模是动态变化的，如何为深度网络的状态输入设计满足动态变化（变长或变维）的状态表示是关键所在，也是当前深度强化学习方法面临的重要挑战。可扩展

展学习主要是降低各种学习方法对输入因素的敏感度, 集合具备置换不变性 (permutation invariance) 的特点, 与集合理论中集合元素置换不变性不同的是, 基于注意力机制的方法更加强调智能体角色、重要性的不同。为了利用多智能体强化学习的组合特性并获得高效和可扩展的算法, 可以使用图来表示智能体之间的稀疏交互。平均场理论是利用平均嵌入的思想计算多个智能体的状态表示平均值, 将变规模智能体的状态编码转换成固定维数张量。未来围绕可扩展学习方法的相关研究可考虑从突破平均场理论极限、建模高阶复杂网络交互、设计满足等变性质的神经网络等角度展开。

3 种群自适应学习方法

DeepMind 发布了关于真实世界博弈策略的陀螺猜想^[33], 实证分析了多类博弈的策略空间满足此猜想, 陀螺型博弈策略空间形态如图9所示。策略之间的传递压制与循环压制并存, 正如现实情景中如何玩好“石头、剪刀、布”一直是个挑战问题, 如何分析各类博弈的策略空间形态一直是一个开放式问题。

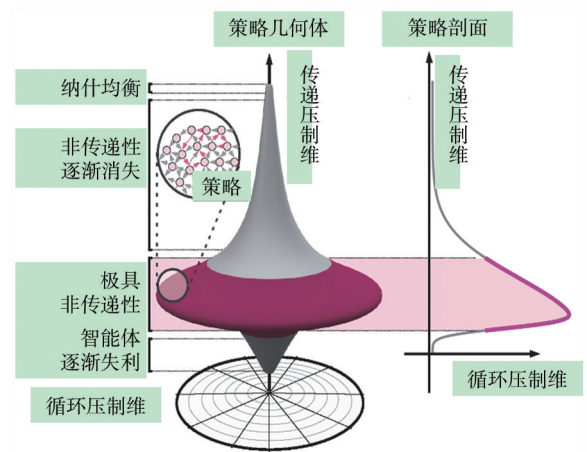


图9 陀螺型博弈策略空间形态^[33]

分布式群智能体强化学习包含两个“群”概念, 即一群智能体和策略种群。需要研究的问题包括两方面: 规模可扩展多智能体强化学习方法, 主要关注如何训练一群智能体; 自适应深度强化学习, 采用基于种群的训练方法, 训练一个智能体种群。种群自适应学习方法主要聚焦于如何通过自适应学习方法得到策略种群, 这是这类方法的本质。当前主流建模框架与典型方法如表2所示。

表2 相关建模框架与典型自适应强化学习方法

类别	建模框架与典型方法	主要机制及方法特点	缺点分析
迁移学习	离线预训练 ^[59]	利用线下大样本, 重构强化学习范式, 进行预训练模式的学习	预训练学习策略需要在线微调, 难处理分布外场景
	任务及域适应 ^[60]	将源任务中学习到的知识, 用于适配目标任务	
	智能体间迁移 ^[61]	智能体之间采用	
课程学习	任务难易程度课程 ^[62]	将不同的任务场景分解成不同难度的子任务	课程难设计, 自主课程中环境-策略协同演化难收敛
	智能体规模课程	设定不同规模数量的智能体学习场景	
	自主课程学习 ^[63]	利用环境与智能体进行协同学习	
元学习	度量元学习 ^[64]	基于深度度量(相似性)学习	训练过程不稳定, 难以适应新的任务
	基础与元学习器 ^[65]	基础学习器学习底层策略, 元学习器学习上层共性策略	
	贝叶斯元学习 ^[66]	从多个模型中推断贝叶斯后验	
元博弈学习	α -Rank 与 PSRO ^[67]	基于策略评估的博弈策略学习方法	大规模博弈策略学习样本效率较低, 分布式并行框架难适配, 学习到的策略模型难以应对高动态测试时规划
	管线 PSRO ^[68]	利用并行化的博弈策略学习方法	
	单纯形 PSRO ^[69]	利用单纯形构建基于贝叶斯最优的策略学习方法	
	自主 PSRO ^[70]	基于课程学习方法设计自主博弈策略学习方法	
	离线 PSRO ^[71]	利用离线资料学习环境模型和预先训练策略模型	
	在线 PSRO ^[72]	考虑对手类型的在线无悔的博弈策略学习方法	
	随时 PSRO ^[73]	基于种群后悔最小化的迭代式博弈策略学习方法	
	自对弈 PSRO ^[74]	基于自对弈学习的迭代式博弈策略学习方法	

3.1 基于迁移学习的方法

迁移学习视角下的智能体策略学习一般可分为两阶段: 源任务中学习或预训练、目标任务中适配。

3.1.1 离线预训练

作为一类试图充分利用离线交互数据的学习范式, 离线预训练通过大量智能体和环境的交互样

本, 利用原来通过强化学习收集的交互模型, 将强化学习过程重构为监督学习过程。如图10所示, 适用于离线预训练的因果Transformer解码器将环境观测、未来奖励、当前行动与当前奖励作为输入^[59]。

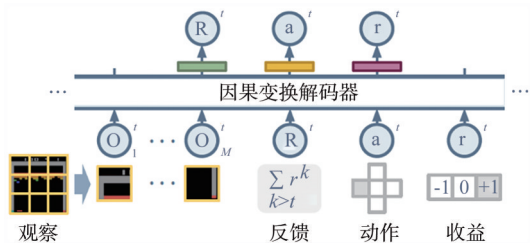


图10 基于因果Transformer解码的离线预训练^[59]

当前, 围绕预训练模型的相关研究主要聚焦在设计更好的预训练范式、提升模型的可扩展性、在线微调与提升等方面。

3.1.2 任务及域适应性

任务与域适应性迁移学习法主要是通过任务关系、任务间相似度、知识经验共享、知识蒸馏和策略复用等手段, 对源任务中多源策略的性能进行评估, 并选择适当的策略进行适配, 从而将源任务中学习到的知识进行迁移。Qin等人^[60]提出基于任务间的关系进行策略迁移。当前, 这类方法仍面临灾难性遗忘、负迁移、参数与数据效率等挑战。

3.1.3 智能体间迁移

智能体之间可以通过策略或知识共享的方式, 提高多智能体场景下的学习效率。当前的一些研究采用教师-学生框架, 点对点教学、行为建议、模仿学习等方式, 策略蒸馏等方法实现智能体之间策略的迁移学习。Zhao等人^[61]提出中心化教师-分散式学生框架, 教师模型通过学习个体Q值来分配团队报酬, 条件是全局性观察, 学生模型则用局部观察值来近似教师模型估计的Q值。

3.2 基于课程学习的方法

课程学习是一套模仿人类先易后难、先简单后复杂的顺序式、渐进性学习方法。基于课程学习的思想, 可以从简单的样本或任务开始, 而后逐步过渡至复杂样本或任务, 从而提高策略学习算法的性能。

3.2.1 任务难易程度课程

针对任务场景的复杂度高的问题, 可以进行子任务分解, 通过奖励塑造、智能体风格偏好等, 设计不同等级难易程度的课程任务。通常采用控制变

量的方式生成一系列不同的课程或将复杂任务拆分成多个子任务。Jia等人^[62]提出可以将智能体学习篮球控制策略的过程分成进攻、防守、助攻、控球等5个子任务。

3.2.2 智能体规模课程

智能体规模数量的不同, 会对学习模型的泛化性能产生影响。传统基于参数共享式的学习方法无法很好地扩展到更多智能体的场景。可以设计围绕数量的分解子课程、动态数量课程、种群进化课程等。从小规模智能体交互场景的策略学习开始, 逐步增加智能体的数量规模。如图11所示, 设计不同数量智能体课程^[47]。

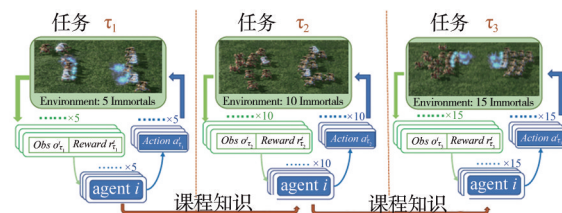


图11 不同数量智能体课程学习^[47]

3.2.3 自主课程学习

自主课程学习方法将智能体与任务环境进行耦合, 迭代优化智能体与环境之间的双向适应能力。给定智能体行为（如性能或访问状态）的度量, 自主课程学习方法生成适应智能体能力的新任务。如图12所示, 自主课程学习可以控制任务的各个元素, 塑造智能体的学习轨迹^[63]。

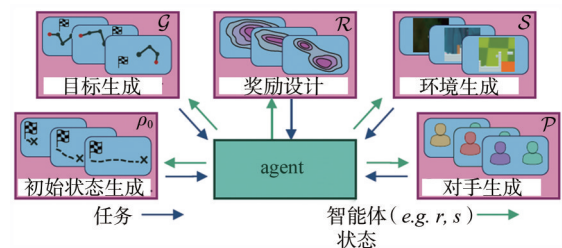


图12 自主课程学习数据收集过程^[63]

3.3 基于元学习的方法

元学习也称学会学习, 可以实现模型的快速准确迁移, 降低模型训练的成本, 让模型快速适应新任务, 特别适合用于小样本和环境不断变化的任务场景的策略学习。现在的元学习法主要有三大类: 度量元学习、基础与元学习器、贝叶斯元学习。

3.3.1 度量元学习

度量元学习方法是深度度量（相似性）学习与

预言机方法利用环境动态模型与对手模型来迭代式学习博弈策略^[71]。

3.4.6 在线 PSRO

离线策略的学习通常采用模拟器将双方的行动策略耦合在一起，在线交互过程中，博弈双方通常处于非耦合状态。在线策略空间响应预言机^[72]通常通过区分对手的类型（随机型、对抗型、遗忘型），分析近似纳什均衡策略的空间形态，采用在线无悔学习类方法，训练一个基于时间后悔值界的策略。

3.4.7 随时 PSRO

基于种群后悔最小化的迭代式博弈策略学习方法，在每轮迭代过程中，学习一个受限的策略分布，每次训练一个相对对手最佳策略具有后悔最小化的应对策略。每一次迭代过程中，当某方处于无约束状态时，创建两个受限博弈，首先使用无悔方法更新一个受限分布，而后训练一个面向约束分布的最佳响应策略^[73]。

3.4.8 自对弈 PSRO

由于 PSRO 类方法在每次迭代过程中，通常只将单个纯策略（确定性）最佳响应添加至种群中。但 PSRO 类方法在找到纳什均衡之前可能会将所有的策略添加至策略种群中。尽管找到要添加的最佳响应与解决最初的博弈一样困难，但在每次迭代时，添加一个近似最佳响应，可以快速减少受限元博弈策略分布的可利用性。基于自对弈 PSRO 方法^[74]，在任意时间策略空间响应预言机方法的基础上，每次迭代采用离策强化学习训练一个新策略，然后往策略池中添加两个策略：时间平均（time-average）新策略和对手受限分布元博弈的最佳响应策略。

3.5 面向自适应性的典型学习平台

EPC^[13]是一类基于演化种群课程学习与注意力机制的扩展性多智能体学习环境。在每个学习阶段维护智能体集合，运用一种课程学习范式，通过循序渐进地增加训练主体的数量来扩大多智能体强化学习的规模。此外，EPC 使用一种演化学习方法。具体来说，EPC 在每个阶段维护多个智能体集合，对这些集合执行混合匹配和微调，形成下一个阶段具有最佳适应性的智能体集合。

MALib^[75]是由上海交通大学设计的基于种群的多智能体深度强化学习（population based MARL，PB-MARL）框架，其本质上是一个基于元博弈理

论设计、具备“策略评估与策略提升”能力的多智能体博弈学习框架。其支持丰富的种群训练方法，如自对弈、策略空间响应预言机（policy space response oracle，PSRO）、联赛训练（league training）。其底层采用 Ray 框架，支撑多类多智能体博弈对抗环境，如星际争霸、谷歌足球、模型类、雅达利游戏和墨子兵棋平台等。

3.6 自适应学习方法小结

自适应学习的任务是根据环境的变化将全局奖励自适应地分配给单个智能体，使得智能体能够据此选择对全局有利的动作，进而提高算法模型的训练效率及性能。迁移学习框架下的智能体自适应策略学习主要分为对源任务的学习或预训练和目标任务的适配两个方面。基于课程学习的方法是模仿人类先易后难、先简单后复杂的顺序式、渐进性的学习方法，该方法的思路是从简单的样本或任务开始，而后逐步过渡至复杂样本或任务。基于元学习的方法可以实现模型的快速准确迁移，降低模型训练的成本，让模型快速适应新任务，特别适合用于小样本和环境不断变化任务场景的策略学习；基于元博弈的策略空间形态理论为基于种群的学习方法设计提供了理论支撑，作为一种迭代式实证博弈理论分析方法，策略空间响应预言机为基于博弈理论的多智能体强化学习提供了统一框架，成为当前主要的学习方法设计范式。未来围绕自适应学习方法的相关研究可考虑从元演化、协同演化、自主课程学习、非对称元博弈学习等角度展开。

4 典型应用与前沿展望

4.1 典型应用场景

4.1.1 自主无人平台控制

随着 5G 通信、云原生和边缘计算、物联网和大数据等新兴技术的更新换代，无人驾驶汽车、无人仓储物流、无人（车、机、艇、船）集群系统等应用日益成熟。无人平台正在逐步取代人类在军事和民用领域从事的枯燥和危险的任务。随着跨学科融合的日益加深，深度融合了电子、通信、人工智能、机械制造等多学科关键技术的无人系统日益多见，并愈发受到关注。无人潜航器、无人水面艇、无人地面车、无人飞行器等以无人技术为主导的新型智能化无人系统正改变着未来社会形态。在民用领域，无人机集群可以应用于农业植保、智能交通、抢险救灾、资源监测等场景；在军用领域，无

人机集群可用于执行通信中继、协同察打、干扰压制、巡逻搜救等任务。规模可扩展多智能体强化学习方法可用于解决这类自主无人平台的控制问题。

4.1.2 资源分配策略优化

决策科学领域的核心问题是资源分配。在资源分配问题中，需要将资源分配给若干目标（对象），并考虑相应的限制条件和需要优化的目标，构建资源分配模型，根据需要分配的资源与目标对象之间的种类和数量得出资源分配方案。在电力分配、网络资源分配、安全设备布设、军事兵力分配、军事物流资源分配、卫星在轨资源分析、云计算、政治选举、频谱资源拍卖、安全巡逻、偷猎走私、兵力分配、战略博弈等领域，很多问题可以被建模成资源分配问题。博弈论与运筹学作为决策科学的两个分支，为资源分配问题提供了建模工具与求解方法。从博弈论视角来看，资源分配可用于多种场景、多类问题的建模。根据博弈局中人与人之间的合作、对抗、混合关系，可将问题分为合作条件下资源分配问题、对抗条件下资源分配问题以及混合式条件下资源分配问题，其中对抗条件下资源分配问题是本文的研究核心。根据博弈局中人与人之间的决策顺序可将博弈建模成 Stackelberg 博弈、正则式博弈、多阶段博弈和马尔可夫博弈。从运筹学的视角来看，面向资源分配的博弈问题可以转化成多重线性规划或凸优化问题，可以利用优化理论和方法来求解，特别是在线资源分配问题。基于种群的自适应强化学习方法可用于优化各类资源分配问题。

4.1.3 分布式智能系统管理

逻辑描述是早期分布式系统形式化建模的主要方式。分布式系统是具有分散或分布式内存和处理资源的节点集，整个系统支持多进程、进程间通信、共享内存等。智能系统可以借助人工智能技术的赋能，利用感知、学习、推理、判断等手段，模拟领域专家解决问题的思维过程，运用大量的领域专家知识和经验，对复杂问题进行有效的处理。近年来，信息物理系统（cyber-physical system, CPS）、物联网（Internet of thing, IoT）等概念和描述模型作为新的分布式智能系统形态诞生了。这些新型系统需要可扩展和适应性管控策略的支撑。

4.2 前沿研究展望

4.2.1 基准平台开发

近年来随着深度强化学习不断取得突破性进展，大量与之相关的算法和应用走入视野，很多研

究不再局限于单智能体强化学习，研究多智能体学习场景下的深度强化学习逐渐成为热点，强化学习中单个智能体与环境交互学习扩展至多智能体在复杂环境中合作竞争与共同进化，在对不同智能体学习策略奖励不同的情况下，不断改进学习算法。目前该方向已经取得了一系列瞩目的进展，比如交通信号控制、机器人控制、未知探索、公交车时刻表优化等，该领域的研究与应用也越来越多，其在通信网络、合作探索等方面都进行了深入的研究。当前部分基准测试环境无法提供可向现实场景迁移的策略，未来可考虑通过构建分布式群智能体学习问题的测试基准平台和算法基准平台解决上述问题。

4.2.2 双层优化架构

当前以双层优化理论^[76]为指引的博弈强化学习、演化强化学习、元学习、持续学习、终身学习等方法为强对抗条件下的智能决策策略学习提供了可供参考的架构。双层优化理论起源于 Stackelberg 博弈（主从博弈）^[77]，当前被广泛应用于多智能体博弈强化学习，Zhang 等人^[78]基于 Stackelberg 博弈提出了面向多智能体协同的双层 Actor-Critic 算法。基于双层优化的思想，李智等人^[79]提出了面向联合全域作战的“智能—博弈—优化”一体化指挥决策框架，尝试利用人工智能、博弈论和多目标优化理论，为联合全域作战智能化指挥决策提供技术支撑。

4.2.3 对抗策略学习

（1）样本效率与分布式：策略探索的样本效率与方法的实用性十分相关，借助分布式计算架构，如何设计满足策略学习的分布式框架与学习方法尤为关键。特别是如何设计满足“智能体—玩家”两层协同演化的分布式架构^[33]。

（2）少样本与多类样本：策略的学习依赖离线或在线样本数据，仿真器失真或在线流式交互样本不足、环境的非平稳、训练样本与测试样本之间的分布迁移等为零样本（zero shot）、少样本（few shot）、多类样本（many shot）学习提出了挑战^[80]。

（3）可扩展与自适应：为了应对博弈参与主体（智能体）数据规模的可变性，需要设计考虑了可扩展性的学习方法^[81]。由于离线（训练）与在线（测试）阶段策略学习的条件不一致，离线时可借助仿真环境将多智能体的交互耦合在一起学习策略，而在线阶段智能体策略与对手策略之间通常为解耦关系，如何调节自适应策略满足在线场景显

得十分重要。

4.2.4 人机协同价值对齐

人机协同已然从初始的人机对抗^[82]、人机共生^[83]走向了人智共生(human-AI symbiosis, HAIS)时代。人工智能系统处理问题的方式通常与人类的方式不同,因此可能无法解释,也很难从中学习。弥合人类和人工智能之间差距的关键一步是对构成人类行为的动作进行建模,而不是简单地匹配人类的总体表现。

虽然超人智能体的决策正确性可能无法评估,但是如果模型的决策不能满足某些合乎逻辑的、人类可解释的规则,我们仍然可以发现错误^[84]。McIlroy等人^[85]基于国际象棋模型,设计了对齐超人AI与人类行为的Maia方法,利用人类国际象棋比赛数据训练,可以更准确地预测人类的“着法”,还可预测人类下一步是否会犯大错误。通过对人类决策行为进行精确建模,可为设计人类协作人工智能系统带来更大可能。

4.2.5 自适应博弈决策环

依赖学习及预训练的人工智能程序在现实中单次适用时,不可避免地要应对分布外(out-of-distribution)情境,回答测试时间适应(test-time adaptation)^[86]、决策时间规划(decision-time planning)^[87]等临机决策问题。

RAND提出了基于“感知-决策-适应-行动”的自适应交战决策模型^[88],如何在基于博弈理论的开放式决策环境中设计适变策略生成模块已然十分重要^[89-90]。

5 结论

传统的多智能体强化学习方法对于少量智能体的应用场景效果比较明显,但对于智能体数据规模过大的场景,指数爆炸导致学习时的样本效率较低。为了应对大规模智能体,设计规模可扩展的多智能体强化学习方法逐渐成为当前研究的焦点。本文围绕多智能体强化学习方法展开综述,概述了多智能体强化学习方法基础,梳理了相关挑战和前沿;分别从规模可扩展和种群自适应两个角度分类梳理了群智能体强化学习方法;最后分析了分布式群智能体强化学习的典型应用,并对未来研究前沿进行了展望,可为研究与多智能系统强化学习有关的重点问题提供参考。

参考文献:

- [1] YANG Y, WANG J. An overview of multi-agent reinforcement learning from game theoretical perspective [EB]. arXiv preprint, 2020, arXiv: 2011.00583.
- [2] SILVA F L, COSTA A H R. A survey on transfer learning for multi-agent reinforcement learning systems[J]. Journal of Artificial Intelligence Research, 2019, 64: 645-703.
- [3] YANG T, TANG H, BAI C, et al. Exploration in deep reinforcement learning: a comprehensive survey [EB]. arXiv preprint, 2021, arXiv: 2109.06668.
- [4] 王涵, 俞扬, 姜远. 基于通信的多智能体强化学习进展综述[J]. 中国科学: 信息科学, 2022, 52(5): 742-764.
WANG H, YU Y, JIANG Y. Review on the progress of multi-agent reinforcement learning based on communication[J]. Science China Information Sciences, 2022, 52(5): 742-764.
- [5] 殷昌盛, 杨若鹏, 朱巍, 等. 多智能体分层强化学习综述 [J]. 智能系统学报, 2020, 15(4): 646-655.
YIN C S, YANG R P, ZHU W, et al. A survey on multi-agent hierarchical reinforcement learning[J]. CAAI Transactions on Intelligent Systems, 2020, 15(4): 646-655.
- [6] 王龙, 黄锋. 多智能体博弈、学习与控制[J]. 自动化学报, 2023, 49(3): 580-613.
WANG L, HUANG F. An interdisciplinary survey of multi-agent games, learning and control [J]. Acta Automatica Sinica, 2023, 49(3): 580-613.
- [7] 罗俊仁, 张万鹏, 苏炯铭, 等. 多智能体博弈学习研究进展[J]. 系统工程与电子技术, 2022, 已录用.
LUO J R, ZHANG W P, SU J M, et al. Research progress of multi-agent game theoretic learning[J]. Systems Engineering and Electronics, 2022, Accepted.
- [8] DENG X, LI N, MGUNI D, et al. On the complexity of computing markov perfect equilibrium in general-sum stochastic games[J]. National Science Review, 2023, 10(1): nwac256.
- [9] ZHANG T, LI Y, LI S, et al. Decentralized circle formation control for fish-like robots in the real-world via reinforcement learning[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2021: 8814-8820.
- [10] ZHENG L, YANG J, CAI H, et al. MAgent: a many-agent reinforcement learning platform for artificial collective intelligence[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1).
- [11] FU Q X, QIU T H, YI J Q, et al. Concentration network for reinforcement learning of large-scale multi-agent systems[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(9): 9341-9349.
- [12] SUAREZ J, DU Y, ZHU C, et al. The neural MMO platform for massively multi-agent research[C]// Proceedings of Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1). [S.l.:s.n.], 2021.
- [13] LONG Q, ZHOU Z, GUPTA A, et al. Evolutionary population curriculum for scaling multi-agent reinforcement learning[C]// Proceedings of International Conference on Learning Representations. Addis Ababa: [s.n.], 2019.
- [14] QU G N, WIERMAN A, LI N. Scalable reinforcement learning for multi-agent networked systems[J]. Operations Research, 2022, 70(6): 3601-3628.
- [15] TAMPUU A, MATIISEN T, KODELJA D, et al. Multi-agent cooperation and competition with deep reinforcement learning[J]. PloS One,

- 2017, 12(4): e0172395.
- [16] LEIBO J Z, PEROLAT J, HUGHES E, et al. Malthusian reinforcement learning[C]//Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems. [S.l.:s.n.], 2019: 1099-1107.
 - [17] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6382-6393.
 - [18] PALMER G, TUYLS K, BLOEMBERGEN D, et al. Lenient multi-agent deep reinforcement learning[C]// Proceedings of AAMAS. International Foundation for Autonomous Agents and Multi-Agent Systems Richland. [S.l.:s.n.], 2018: 443-451.
 - [19] FOERSTER J, FARQUHAR G, AFOURAS T, et al. Counterfactual multi-agent policy gradients[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 2974-2982.
 - [20] SUNEHAG P, LEVER G, GRUSLYS A, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward[C]// Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems. New York: ACM, 2018: 2085-2087.
 - [21] RASHID T, SAMVELYAN M, SCHROEDER C, et al. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning[C]//Proceedings of International Conference on Machine Learning. New York: PMLR, 2018: 4295-4304.
 - [22] FOERSTER J N, ASSAEL Y M, DE FREITAS N, et al. Learning to communicate with deep multi-agent reinforcement learning[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM, 2016: 2145-2153.
 - [23] UKHBAATAR S, SZLAM A, FERGUS R. Learning multi-agent communication with backpropagation[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM, 2016: 2252-2260.
 - [24] PENG P, WEN Y, YANG Y, et al. Multi-agent bidirectionally-coordinated nets: emergence of human-level coordination in learning to play starcraft combat games [EB]. arXiv preprint, arXiv:1703.10069.
 - [25] HE H, BOYD-GRABER J, KWOK K, et al. Opponent modeling in deep reinforcement learning[C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. New York: ACM, 2016: 1804-1813.
 - [26] FOERSTER J, CHEN R Y, AL-SHEDIVAT M, et al. Learning with opponent-learning awareness[C]//Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems. New York: ACM, 2018: 122-130.
 - [27] RABINOWITZ N, PERBET F, SONG F, et al. Machine theory of mind[C]//Proceedings of International Conference on Machine Learning. New York: PMLR, 2018: 4218-4227.
 - [28] LI S H, WU Y, CUI X Y, et al. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 4213-4220.
 - [29] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6382-6393.
 - [30] MAO H Y, ZHANG Z C, XIAO Z, et al. Modelling the dynamic joint policy of teammates with attention multi-agent DDPG[C]//Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems. New York: ACM, 2019: 1108-1116.
 - [31] 熊丽琴, 曹雷, 赖俊, 等. 基于值分解的多智能体深度强化学习综述[J]. 计算机科学, 2022, 49(9): 172-182.
 - XIONG L Q, CAO L, LAI J, et al. Overview of multi-agent deep reinforcement learning based on value factorization[J]. Computer Science, 2022, 49(9): 172-182.
 - [32] GUESTRIN C, KOLLER D, PARR R. Multi-agent planning with factored MDPs[C]// Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. [S.l.:s.n.], 2001: 1523-1530.
 - [33] CZARNECKI W M, GIDEL G, TRACEY B, et al. Real world games look like spinning tops[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 17443-17454.
 - [34] YIN Q, YU T, SHEN S, et al. Distributed deep reinforcement learning: a survey and a multi-player multi-agent learning toolbox [EB]. arXiv preprint, 2022, arXiv:2212.00253v1.
 - [35] ZAHEER M, KOTTUR S, RAVANBHAHSH S, et al. Deep sets[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. [S.l.:s.n.], 2017: 3394-3404.
 - [36] HUEGLE M, KALWEIT G, MIRCHEVSKA B, et al. Dynamic input for deep reinforcement learning in autonomous driving[C]// IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE, 2019: 7566-7573.
 - [37] SHI G Y, HÖNIG W, YUE Y S, et al. Neural-swarm: decentralized close-proximity multirotor control using learned interactions[C]//Proceedings of 2020 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2020: 3241-3247.
 - [38] SHI G, HÖNIG W, SHI X, et al. Neural-swarm2: planning and control of heterogeneous multirotor swarms using learned interactions[J]. IEEE Transactions on Robotics, 2021, 38(2): 3098436.
 - [39] LI Y, WANG L, YANG J, et al. Permutation invariant policy optimization for mean-field multi-agent reinforcement learning: a principled approach [EB]. arXiv preprint, 2021, arXiv:2105.08268.
 - [40] LIU X, TAN Y. Attentive relational state representation in decentralized multi-agent reinforcement learning[J]. IEEE Transactions on Cybernetics, 2022, 52(1): 252-264.
 - [41] IQBAL S, WITT C D, PENG B, et al. Randomized entity-wise factorization for multi-agent reinforcement learning[C]// Proceedings of International Conference on Machine Learning (ICML). New York: PMLR, 2021: 4596-4606.
 - [42] BATRA S, HUANG Z, PETRENKO A, et al. Decentralized control of quadrotor swarms with end-to-end deep reinforcement learning[EB]. arXiv preprint, 2021, arXiv:2109.07735.
 - [43] HU S, ZHU F, CHANG X, et al. UPDeT: universal multi-agent reinforcement learning via policy decoupling with transformers[C]// Proceedings of International Conference on Learning Representations (ICLR). [S.l.:s.n.], 2021.
 - [44] ZHOU T, ZHANG F, SHAO K, et al. Cooperative multi-agent transfer learning with level-adaptive credit assignment [EB]. arXiv preprint, 2021, arXiv:2106.00517.
 - [45] NAIR R, VARAKANTHAM P, TAMBE M, et al. Networked distributed POMDPs: a synthesis of distributed constraint optimization and POMDPs[C]// Proceedings of the 20th National Conference on Artificial Intelligence-Volume 1. New York: ACM, 2005: 133-139.
 - [46] OLIEHOEK F A, WHITESON S, SPAAN M T J. Approximate solutions for factored dec-POMDPs with many agents[C]// Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems. New York: ACM, 2013: 563-570.

- [47] LIU I J, YE H R A, Schwing A G. PIC: permutation invariant critic for multi-agent deep reinforcement learning[C]// Conference on Robot Learning (CoRL). New York: PMLR, 2020: 590-602.
- [48] WANG W, YANG T, LIU Y, et al. From few to more: large-scale dynamic multi-agent curriculum learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 7293-7300.
- [49] NADERIALIZADEH N, HUNG F H, Soleyman S, et al. Graph convolutional value decomposition in multi-agent reinforcement learning [EB]. arXiv preprint, 2020, arXiv:2010.04740.
- [50] RYU H, SHIN H, PARK J. Multi-agent actor-critic with hierarchical graph attention network[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 7236-7243.
- [51] YE Z H, WANG K, CHEN Y N, et al. Multi-UAV navigation for partially observable communication coverage by graph reinforcement learning[J]. IEEE Transactions on Mobile Computing, 2023, 22(7): 4056-4069.
- [52] BÖHMER W, KURIN V, WHITESON S. Deep coordination graphs [C]// Proceedings of International Conference on Machine Learning. New York: PMLR, 2020: 980-991.
- [53] BAI Y, GONG C, ZHANG B, et al. Value function factorisation with hypergraph convolution for cooperative multi-agent reinforcement learning [EB]. arXiv preprint, 2021, arXiv:2112.06771.
- [54] OLIEHOEK F A, SPAAN M T J, WHITESON S, et al. Exploiting locality of interaction in factored dec-POMDPs[C]// Proceedings of the 7th International Joint Conference on Autonomous Agents and Multi-Agent Systems. [S.l.:s.n.], 2008: 517-524.
- [55] YANG Y, LUO R, LI M, et al. Mean field multi-agent reinforcement learning[C]// Proceedings of International Conference on Machine Learning (ICML). New York: PMLR, 2018: 5571-5580.
- [56] LAURIÈRE M, PERRIN S, GIRGIN S, et al. Scalable deep reinforcement learning algorithms for mean field games[C]// Proceedings of the Thirty-Ninth International Conference on Machine Learning. [S.l.:s.n.], 2022:12078-12095.
- [57] CAINES P E, HUANG M Y. Graphon mean field games and the GMFG equations: ϵ -Nash equilibria[C]//Proceedings of 2019 IEEE 58th Conference on Decision and Control (CDC). Piscataway: IEEE Press, 2019: 286-292.
- [58] GU H, GUO X, WEI X, et al. Mean-field controls with Q-learning for cooperative MARL: convergence and complexity analysis[J]. SIAM Journal on Mathematics of Data Science, 2021, 3(4): 1168-1196.
- [59] LEE K H, NACHUM O, YANG M, et al. Multi-game decision transformers [EB]. arXiv preprint, 2022, arXiv:2205.15241.
- [60] QIN R, CHEN F, WANG T, et al. Multi-agent policy transfer via task relationship modeling [J]. arXiv preprint, 2022, arXiv:2203.04482.
- [61] ZHAO J, HU X, YANG M, et al. CTDS: centralized teacher with decentralized student for multi-agent reinforcement learning [EB]. arXiv preprint, 2022, arXiv:2203.08412.
- [62] JIA H T, REN C X, HU Y J, et al. Mastering basketball with deep reinforcement learning: an integrated curriculum training approach[C]// Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems. New York: ACM, 2020: 1872-1874.
- [63] PORTELAS R, COLAS C, WENG L, et al. Automatic curriculum learning for deep RL: a short survey[C]// Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. [S.l.:s.n.], 2021: 4819-4825.
- [64] LI L, YANG R, LUO D. FOCAL: efficient fully-offline meta-reinforcement learning via distance metric learning and behavior regularization[C]// Proceedings of International Conference on Learning Representations. [S.l.:s.n.], 2020.
- [65] SHAO J, ZHANG H, JIANG Y, et al. Credit assignment with meta-policy gradient for multi-agent reinforcement learning [EB]. arXiv preprint, 2021, arXiv:2102.12957.
- [66] YOON J, KIM T, DIA O, et al. Bayesian model-agnostic meta-learning[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. [S.l.:s.n.], 2018: 7343-7353.
- [67] MULLER P, OMIDSHAFIEI S, ROWLAND M, et al. A generalized training approach for multi-agent learning[C]// Proceedings of the 8th International Conference on Learning Representations. [S. l.: s. n.], 2020: 1-35.
- [68] MCALEER S, LANIER J, FOX R, et al. Pipeline PSRO: a scalable approach for finding approximate Nash equilibria in large games[C]// Proceedings of Advances in Neural Information Processing Systems. [S.l.:s.n.], 2020, 33: 20238-20248.
- [69] LIU S, LANCTOT M, MARRIS L, et al. Simplex neural population learning: any-mixture bayes-optimality in symmetric zero-sum games [EB]. arXiv preprint, 2022, arXiv:2205.15879.
- [70] FENG X, SLUMBERS O, YANG Y, et al. Neural auto-curricula in two-player zero-sum games[C]// Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems. [S.l.:s.n.], 2021: 3504-3517.
- [71] LI S, WANG X, CERNY J, et al. Offline equilibrium finding [EB]. arXiv preprint, 2022, arXiv:2207.05285.
- [72] DINH L C, YANG Y, TIAN Z, et al. Online double oracle [EB]. arXiv preprint, 2021, arXiv:2103.07780.
- [73] MCALEER S, WANG K, LANCTOT M, et al. Anytime optimal PSRO for two-player zero-sum games [EB]. arXiv preprint, 2022, arXiv:2201.07700.
- [74] MCALEER S, LANIER J B, WANG K, et al. Self-play PSRO: toward optimal populations in two-player zero-sum games [EB]. arXiv preprint, 2022, arXiv:2207.06541.
- [75] ZHOU M, WAN Z, WANG H, et al. Malib: a parallel framework for population-based multi-agent reinforcement learning [EB]. arXiv preprint, 2021, arXiv:2106.07551.
- [76] SINHA A, MALO P, DEB K. A review on bilevel optimization: from classical to evolutionary approaches and applications[J]. IEEE Transactions on Evolutionary Computation, 2017, 22(2): 276-295.
- [77] CHEN L S, JOSE S T, NIKOLOSKA I, et al. Learning with limited samples: meta-learning and applications to communication systems[J]. Foundations and Trends® in Signal Processing, 2023, 17(2): 79-208.
- [78] ZHANG H F, CHEN W Z, HUANG Z R, et al. Bi-level actor-critic for multi-agent coordination[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 7325-7332.
- [79] 李智, 吕铁鑫, 潘艳辉. 联合全域作战智能博弈优化一体化决策问题[J]. 火力与指挥控制, 2023, 48(3):1-8.
- LI Z, LYU T X, PAN Y H. Research on integrated decision-making problems of intelligent game optimization in JADO[J]. Fire Control & Command Control, 2023, 48(3):1-8.
- [80] BECK J, VUORIO R, LIU E Z, et al. A survey of meta-reinforcement learning [EB]. arXiv preprint, 2023, arXiv:2301.08028.
- [81] CUI K, TAHIR A, EKINCI G, et al. A survey on large-population systems and scalable multi-agent reinforcement learning [EB]. arXiv preprint, 2022, arXiv:2209.03859.
- [82] 黄凯奇, 兴军亮, 张俊格, 等. 人机对抗智能技术[J]. 中国科学: 信息科学, 2020, 50(4): 540-550.
- HUANG K Q, XING J L, ZHANG J G, et al. Intelligent technology of man-machine confrontation [J]. Science China Information Sciences, 2020, 50(4): 540-550.

- [83] GRAYSON T P, LILLIU S. Mosaic warfare and human - machine symbiosis[J]. Scientific Video Protocols, 2021, 1(1): 1-12.
- [84] FLURI L, PALEKA D, TRAMÈR F. Evaluating superhuman models with consistency checks [EB]. arXiv preprint, 2023, arXiv:2306.09983.
- [85] MCILROY-YOUNG R, SEN S, KLEINBERG J, et al. Aligning superhuman AI with human behavior: chess as a model system[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2020: 1677-1687.
- [86] LIANG J, HE R, TAN T. A comprehensive survey on test-time adaptation under distribution shifts [EB]. arXiv preprint, 2023, arXiv:2303.15361.
- [87] SOKOTA S, FARINA G, WU D J, et al. The update equivalence framework for decision-time planning [EB]. arXiv preprint, 2023, arXiv:2304.13138.
- [88] RAND Corporation. Adaptive engagement for undergoverned spaces: concepts, challenges and prospects for new approaches[R]. 2022.
- [89] TEAM A A, BAUER J, BAUMLI K, et al. Human-timescale adaptation in an open-ended task space [EB]. arXiv preprint, 2023, arXiv:2301.07608.
- [90] 王涵,俞扬,姜远. 基于动态自选择参数共享的合作多智能体强化学习算法[J]. 智能科学与技术学报, 2022, 4(1): 75-83.
- WANG H, YU Y, JIANG Y. A cooperative multi-agent reinforcement learning algorithm based on dynamic self-selection parameters sharing [J]. Chinese Journal of Intelligent Science and Technology, 2022, 4(1): 75-83.

[作者简介]



项凤涛 (1986-), 男, 博士, 国防科技大学智能科学学院副教授, 主要研究方向为智能辅助决策、不确定性推理、智能控制。



罗俊仁 (1989-), 男, 国防科技大学智能科学学院博士生, 主要研究方向为不完美信息博弈、多智能体学习等。



谷学强 (1983-), 男, 博士, 国防科技大学智能科学学院副研究员, 主要研究方向为智能规划与决策、智能控制。



苏炯铭 (1984-), 男, 博士, 国防科技大学智能科学学院副研究员, 主要研究方向为可解释人工智能、智能博弈。



张万鹏 (1981-), 男, 博士, 国防科技大学智能科学学院研究员、博士生导师, 主要研究方向为大数据智能、智能演进等。