计算机科学
COMPUTER SCIENCE

面向人机协作的智能体训练方法研究综述

黄炜烨, 陈希亮, 赖俊

引用本文

黄炜烨, 陈希亮, 赖俊. 面向人机协作的智能体训练方法研究综述[J]. 计算机科学, 2025, 52(10): 176-189.

HUANG Weiye, CHEN Xiliang, LAI Jun. Review of Research on Agent Training Methods Toward Human-Agent Collaboration [J]. Computer Science, 2025, 52(10): 176-189.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241000047

面向人机协作的智能体训练方法研究综述

黄炜烨　陈希亮　赖　俊
陆军工程大学指挥控制工程学院　南京 210007
(hwy1115@qq.com)

摘　要　人机协作近年来受到广泛关注,多智能体强化学习在人机协作领域展现出了显著的优势和应用潜力。首先,对多智能体强化学习的基本概念和重要模型进行了介绍,分析了多智能体强化学习在人机协作任务中的优势,并将人机协作分为3种类型进行介绍。其次,论述了多智能体强化学习的3种训练范式,包括集中训练集中执行、分散训练分散执行和集中训练分散执行,以及每种训练范式的适用场景。接着,针对人机协作中智能体训练方法存在的泛化能力差、训练伙伴缺乏多样性以及无法更好地适应人类合作伙伴等问题,从是否使用人类数据的角度,论述了面向人机协作的智能体训练方法的研究进展。最后,讨论了人机协作的应用场景和未来发展趋势,提出了可能的解决思路与研究方向。

关键词:人工智能;多智能体强化学习;人机协作;零样本协调
中图分类号　TP181

Review of Research on Agent Training Methods Toward Human-Agent Collaboration

HUANG Weiye,CHEN Xiliang and LAI Jun
College of Command and Control Engineering,Army Engineering University of PLA,Nanjing 210007,China

Abstract　Human-agent collaboration has received widespread attention in recent years,and multi-agent reinforcement learning has demonstrated significant advantages and application potential in the field of human-agent collaboration. This paper first introduces the basic concepts and important models of multi-agent reinforcement learning,and analyzes the advantages of multi-agent reinforcement learning in human-agent collaborative tasks,and introduces human-agent collaboration in three types. Secondly,it explores three training paradigms of multi-agent reinforcement learning,including centralized training and centralized execution,decentralized training and decentralized execution,and centralized training and decentralized execution,as well as the applicable scenarios for each training paradigm. Then,in response to the problems faced by agent training methods for human-agent collaboration,such as poor generalization ability,lack of diversity in training partners and inability to better adapt to human partners,it summarizes the research progress on agent training methods for human-agent collaboration from the perspective of whether human data is used or not. Finally,it discusses the application scenarios and future development trends of human-agent collaboration,proposes possible solutions and research directions.

Keywords　Artificial intelligence,Multi-agent reinforcement learning,Human-agent collaboration,Zero-shot coordination

1　引言

随着人工智能技术研究的不断深入,人机协作(Human-Agent Collaboration,HAC)[1]技术也得到了广泛的关注和使用。人机协作不仅是技术进步的重要体现,也是推动社会经济发展、提升工作效率、增强创新能力以及实现可持续发展的关键力量,并且随着技术的不断进步和应用场景的拓展,人机协作的未来将更加广阔和深入。

在强化学习的范式中,智能体与环境交互,并根据从环境中收到的奖励或惩罚不断优化其策略。强化学习的早期工作依赖于人工提取的特征,将其输入到线性模型中进行值估计和近似,在复杂场景中表现不佳。

同时,许多现实世界的问题被证明是大规模的、复杂的、实时的且不确定的,将此类问题建模为单智能体系统既低效又不符合实际情况,因此将其建模为多智能体系统(Multi-agent System,MAS)[2]问题更合适。以自动驾驶为例,在交通系统中,单个自动驾驶汽车可以被视为一个单智能体系统;但当考虑整个交通网络时,每辆汽车、行人、交通信号灯等都需要相互协调和通信,这时就需要将整个交通系统建模为多智能体系统,并对整个交通网络中的实体进行统一管理[3]。

人机协作智能体的训练过程通常分为两个步骤。1)获得一个指定策略的单智能体或者多样化策略的智能体样本池;

到稿日期:2024-10-11　返修日期:2025-02-05
基金项目:国家自然科学基金(62273356)
This work was supported by the National Natural Science Foundation of China(62273356).
通信作者:陈希亮(383618393@qq.com)

2)使用生成的样本来训练一个智能体,以进行协作任务。

针对第一个步骤,在人机协作的智能体训练中存在诸多问题,如协作智能体泛化能力差,训练过程中怎样获得多样性训练伙伴,如何更好地适应人类合作伙伴的偏好和期望,数据利用率低,大量使用人类数据样本导致人类隐私泄露等。

最初的方法大多使用大量人类数据样本来训练智能体,如行为克隆博弈(BCP)[4-5],但这又会引起另外的问题,即智能体在训练过程中使用大量人类数据会使训练成本显著增加,且导致人类隐私泄露等问题。若不使用人类数据样本,则大多数标准的多智能体强化学习技术,如自我博弈(Self-play,SP)或群体博弈(Population-play,PP),产生的智能体过于适合其训练伙伴,并且不能很好地推广到不同的人类伙伴。于是,研究者提出了基于零样本协调(ZSC)的方法。例如DeepMind团队提出的虚拟协同博弈(FCP)[6],在不使用人类数据的前提下,增强了智能体的泛化能力。但在不使用人类数据的情况下,智能体可能无法适应人类合作伙伴的偏好。为更好地解决上述问题,研究者提出了使用少样本(Few-shot)训练智能体的方法,如少样本协调(FSC),其在Hanabi上取得了较好的实验结果[7]。

针对第二个步骤,在人机协作智能体的训练过程中,环境往往是复杂且不确定的,而多智能体强化学习能够处理这种不确定性。因此,在人机协作领域,多智能体强化学习(Multi-agent Reinforcement Learning,MARL)为建模和解决上述问题提供了强有力的支持。在多智能体强化学习中,训练范式可以被分为集中训练集中执行(CTCE)、分散训练分散执行(DTDE)以及集中训练分散执行(CTDE)[8]。多智能体强化学习训练范式会直接影响人机协作中智能体训练结果,因此,在人机协作智能体的训练过程中需要考虑使用哪种训练范式,以获得更好的训练效果。

本文首先分类分析了人机协作的不同方式,并介绍了多智能体强化学习的3种训练范式;然后从是否使用人类数据样本的角度进行分类,梳理了近年来主流的人机协作智能体训练方法,并分析了不同方法的特点以及存在的问题;最后对人机协作的现状和未来的发展进行了综述分析。

2 基本概念

本章主要综述了多智能体强化学习和人机协作的基本概念,其中包括强化学习和集中建模方法的基本概念,如MDP、POMDP以及人机协作问题的建模方法;并且从人和智能体在人机协作中所处的地位和所做的工作角度,对人机协作的类型进行了分类分析。

2.1 强化学习

强化学习(Reinforcement Learning,RL)[9]属于机器学习的一个分支,是解决序列决策问题的有效方法[10]。强化学习的目标是最大化累积奖励,智能体必须通过反复试验来探索和学习,以找到最佳策略。该过程可以建模为马尔可夫决策过程(Markov Decision Process,MDP)[10]。MDP是描述智能体如何在完全可观测的环境中进行决策的数学模型[11]。

多智能体系统由分布式人工智能(Distributed Artificial Intelligence,DAI)发展而来,有着高效、低成本、灵活性和可靠性的特点,用于解决使用单个智能体无法高效率建模的问题[2]。根据任务的性质,MAS被分为3种:完全合作型、完全竞争型和混合型。在完全合作的环境中,智能体必须考虑队友的策略,以实现最佳的协调,从而达到奖励最大值。在完全竞争的环境中,智能体之间存在竞争关系,它们的目标是相互对立的。在混合型的环境中,智能体之间通常是合作的,但在一些情况下,例如训练过程中,智能体之间可能需要通过对抗竞争来优化其策略。

多智能体强化学习[12]是强化学习技术在MAS领域的应用。MARL的目标是使用RL方法在共享环境中训练多个智能体来完成指定的任务[13]。当执行协作任务,并且在部分可观测的环境中时,与单智能体强化学习(Single-agent Reinforcement Learning,SARL)中将多步决策过程建模为部分可观测马尔可夫决策过程(Partially Observable Markov Decision Process,POMDP)不同,MARL通常被建模为去中心化部分可观测马尔可夫决策过程(Dec-POMDP)[14-15]。

多智能体强化学习的基本框架如图1所示。在每个时间步 $1\leqslant t\leqslant T$,智能体 $i\in n$ 处于状态 $s_i\in S_T$,根据所处状态,智能体选择动作 $a_i\in A$,所有智能体选择的动作形成联合动作 $A_t=(a_1,\cdots,a_n)$;随后执行该联合动作,并返回给环境。环境接收到智能体的动作后,根据联合动作转换到下一个状态 $s'\sim P(\cdot\mid s,a)$,其中 $s'=S_{t+1}$,$s=S_t$,$a=A_t$。接着,环境返回给智能体 i 它自己的奖励 $R_i(s_i,a_i)$。每个智能体优化其策略函数 $\pi_i:S\to\Delta(A_i)$,以最大化其预期累积奖励。

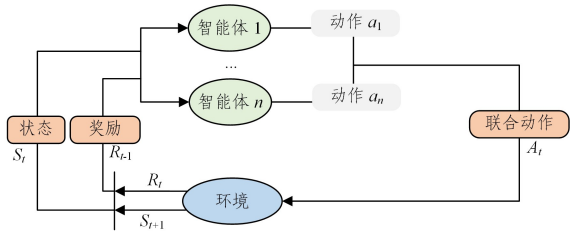


图1 多智能体强化学习基本结构

Fig. 1 Basic structure of multi-agent reinforcement learning

与SARL中智能体只需要考虑自身对环境的影响不同,在MARL中,智能体相互影响,共同做出决策,同时更新策略。

针对人机协作任务的性质,可以将人机协作问题建模为POMDP的扩展。该任务模型可以被表述为一个元组 $\langle N,H,S,A^N,A^H,O,M,r,P,\gamma\rangle$[1]。其中,$N$ 和 H 分别代表智能体和人类的数量,S 是全局状态空间,$A^N=\{A_i^N\}_{i=1,\cdots,N}$ 和 $A^H=\{A_i^H\}_{i=1,\cdots,H}$ 分别表示 N 个智能体和 H 个人类的动作空间,$O=\{O_i\}_{i=1,\cdots,N+H}$ 表示 N 个智能体和 H 个人类的观察空间,M 表示可解释消息的空间,P 和 r 分别表示 N 个智能体的共享状态转移概率函数和奖励函数,γ 为折扣因子。

2.2 人机协作

人机协作是人类与具备自主决策能力或智能处理能力的智能体之间的合作与协同,旨在共同完成特定任务或解决问题[1]。人类与人工智能(智能体)协作的概念与人类人工智能交互(HAI)[16]或人类与机器人交互(HRI)有关[17]。HAC不仅限于简单的任务分配和执行,更强调双方之间的信息交换、

决策制定和协同工作。在 HAC 中，为了达到有效帮助人类所需的智能水平，智能体需要具有社交感知能力，即理解人类行为的能力；以及协作规划的能力，即推理环境并规划其行动以与人类协调[18]。人机交互的概念侧重于人类与 AI 系统之间的直接互动过程。在人机协作中，人机交互主要指合作中的人类参与者与 AI 智能体之间的互动，如收集人类对不同智能体的偏好数据，并将这些数据应用到后续智能体的策略优化中[6]。

此外，HAC 将人类的智能与机器的智能相结合，实现优势互补。例如，人类擅长创造性思维、复杂决策和情感理解，而智能体擅长数据处理、快速响应和精确执行；智能体可以承担繁琐、重复的工作，而人类则可以专注于创造性、战略性的任务。因此，人机协作可以大大提高工作效率和准确性[19]。

实现智能体和人类之间的高效协作，一直是人工智能的长期目标。当下 HAC 的研究目的是增强人类参与者和智能体之间的协作以完成特定的任务。鉴于人机协作过程需要考虑不同场景下的人类和智能体的角色分配、权力关系以及各自的优势和局限性等问题，可以从人类和智能体在协作过程中的不同地位和所做的工作角度将人机协作的方式分为 3 类：人类主导型协作、智能体自主型协作和人机平等型协作[20-22]。

2.2.1 人类主导型协作

在人类主导的人机协作中，人类处于决策的核心地位，并因其具有独特的创造力、判断力和理解力而负责决策制定、战略规划等高级任务[23]。在这种协作下，人类能够直接控制任务的整体方向和进程，确保任务按照预定目标进行，同时能够根据实时情况调整策略和指令来适应环境或任务需求的变化。而智能体则作为高效的执行工具为人类提供辅助性工作，以其强大的数据处理能力、精确的执行能力和持续的工作能力，极大地减轻了人类的负担，提升了整体的工作效率[24]；并且智能体能够迅速处理大量信息并提取出关键数据，为人类的决策提供有力支持。

这种协作模式成功的关键在于，人类能够清晰地定义任务需求，为智能体提供明确的策略或指令；而智能体则通过反馈机制，实时向人类报告任务进度、遇到的问题及潜在的解决方案，使人类能够迅速地做出反应，进一步优化策略。

在这种协作类型中，智能体并不具备自主决策能力，而仅仅具备一定的智能处理能力。

2.2.2 智能体自主型协作

在智能体自主型协作方式下，智能体的角色更为突出，它不再是协作中的辅助角色，但是人类仍然是最终决策者[25]。因此在智能体自主型的协作场景中，智能体不仅具备强大的智能处理能力，也具备自主决策能力，甚至能够主动向人类伙伴提供策略建议。这就要求智能体具备较高的智能水平和自主能力，即自主智能体，以便更有效地与人类协作[26]。

自主智能体在人机协作中的优势不仅包括高效执行任务、持续稳定运行等，更重要的是它们能在一定程度上减轻人类在高风险工作中的负担，降低潜在危险[27]。然而，人类对最终决策权和对任务整体方向的把握仍然是不可或缺的，因此信任问题是这种协作方式中的重要问题。人机协作中，人和智能体建立高度的恰当信任关系，对人和智能体之间的角色分配、工作划分等有很大的影响，恰当的职责划分有助于人和智能体明确彼此的行为策略，更好地协调彼此的动作，更有效地完成协作任务[28]。

2.2.3 人机平等型协作

人机平等型的协作方式作为人机协作的一种更高级形式，改变了以往人类对于“人机互动”的认知边界，人类与智能体不再是传统的主从或辅助协作关系，而是转变为深度互动、相互依赖的平等合作伙伴关系[23]。这种协作方式下，人机共同承担责任，相互合作完成任务。这种平等体现在决策权、任务分配、结果评估等各个方面，人类和智能体通过紧密的合作来实现奖励最大化[29]。

这种方式下的 HAC 问题可以被分为两种场景：Human-to-Agent(H2A)和 Agent-to-Human(A2H)[1]。这不仅体现了人机平等协作下的信息共享，更展示了人类与智能体在决策过程中的灵活切换与互补优势。在 H2A 场景中，人类的直觉、判断力和创造力为智能体提供了方向性的指导；而智能体则以其强大的数据处理能力和策略优化能力，根据人类的宏观期望选择最佳策略与人类协作。反之，在 A2H 场景中，智能体凭借其更深入的数据分析和预测能力，为人类提供策略建议；人类则运用自身的经验和判断，根据智能体的价值体系选择最佳的策略与智能体进行协作。这两种场景的目标都是人类和智能体通过预定义的通信协议进行宏观策略通信，然后选择最佳策略进行有效合作，以获得奖励最大化。

此外，人机平等型协作系统需要具备较高的安全性和可靠性。系统须建立全面的风险评估与防控体系，利用监测技术和智能预警机制，及时发现并有效应对潜在的安全威胁。同时，加强人机之间的沟通与理解，提升人机之间的信任，也是确保系统稳定运行的关键。

人机平等型协作标志着未来人机关系的发展方向，它不仅要求在技术层面不断创新与突破，更需要在理念层面为人类与智能体树立平等、互信、共赢的价值观。

3 种类型的人机协作对比如表 1 所列。

表 1 3 种类型的人机协作的对比

Table 1 Comparison of three types of human-agent collaboration

类型	定义	特点
人类主导型协作	人类处于决策的核心地位，负责决策制定、战略规划等高级任务；智能体作为高效的执行工具，为人类提供辅助性工作	人类直接控制任务的整体方向和进程； 智能体并不具备自主决策能力，仅具备一定的智能处理能力
智能体自主型协作	智能体并不真正处于主导地位，仅仅是其角色更为突出，人类仍然是最终决策者	智能体不仅具备智能处理能力，也具备自主决策能力
人机平等型协作	人类与智能体不再是传统的主从或辅助协作关系，而是转变为深度互动、相互依赖的平等合作伙伴关系	人机共同承担责任，相互合作完成任务

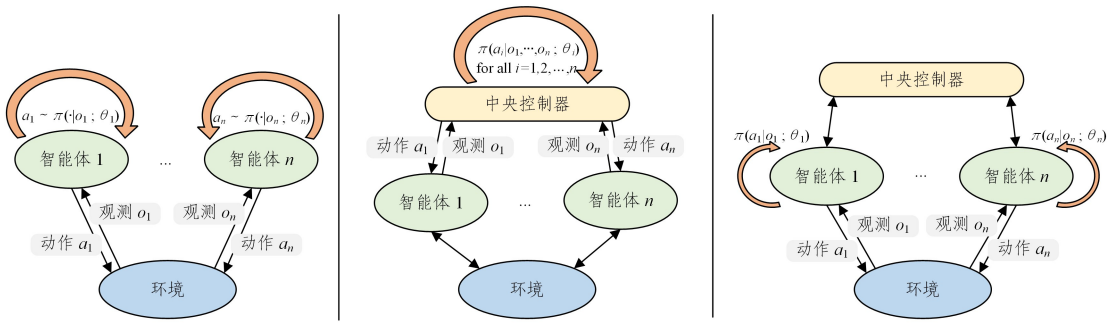
3 多智能体强化学习训练范式

上一章探讨了多智能体强化学习以及人机协作的基本概念，为进一步探讨 MARL 在人机协作智能体训练中的应用提供了框架，并理解智能体如何在复杂的环境中进行协作。基于这些理论基础，本章将深入探讨 MARL 如何具体应用在人机协作领域。

在人机协作领域，多智能体强化学习训练范式显著影响智能体性能。其核心优势在于增强协作能力、泛化鲁棒性并促进持续学习。

在 MARL 中，智能体训练过程主要聚焦于通过积累经验（包括状态、采取的动作以及获得的奖励等）来优化每个智能体的策略[11]。而执行则是指智能体依据其个体策略或联合策略选择并执行动作来实现与环境交互的过程。根据智能体在策略更新时是否需要其他智能体的信息，训练过程可以被分为两类：集中训练和分散训练。集中训练允许智能体在优化其策略时利用来自其他智能体的全局信息，从而提升协作效率与一致性；而分散训练则要求智能体仅基于自身的局部观察进行策略优化，赋予智能体独立决策权，基于局部观测灵活应对环境变化，增强系统鲁棒性。相应地，根据智能体是否需要外部信息或中心控制，执行阶段可以分为集中执行和分散执行。

结合这两个阶段，MARL 包含了 3 种主要范式：集中训练集中执行（CTCE）、分散训练分散执行（DTDE）和集中训练分散执行（CTDE）[8]。这 3 种范式的结构如图 2 所示。



(a)分数训练分散执行 (b)集中训练集中执行 (c)集中训练分散执行

图 2 3 种 MARL 训练范式的结构图

Fig. 2 Structure diagrams of three MARL training paradigm

由于不同的训练方式和执行方式会对人机协作智能体的训练结果产生不同的效果，因此在人机协作中需要考虑采用何种多智能体强化学习训练范式来训练智能体，以适应协作关系，完成协作任务。

3.1 分散训练分散执行

如图 2(a)所示，在 DTDE 框架中，每个智能体 $i \in n$ 独立与环境进行交互，获得状态 s_i 以及所得奖励 r_i，然后根据状态 s_i 和奖励 r_i 独立地做出决策 a_i 并执行，并且每个智能体仅使用其本地信息更新自己的策略。该策略可以表示为 $\pi_i: \Omega_i \rightarrow \Delta(A_i)$。

DTDE 框架具有很好的鲁棒性。在异质环境中时，智能体之间可能缺乏共识，因此允许智能体仅使用其本地信息独立更新策略可以使 MAS 更好地适应环境的不确定性和动态变化。Wu 等[30]在自动驾驶的场景下，基于 DTDE 框架提出 iPLAN 方法，以更好地适应复杂、异质的 MAS 环境，如异质高速公路场景，并提高智能体在决策过程中的自主性和适应性。在非合作导航和异构高速公路两个模拟环境的实验中，iPLAN 的成功率得到了明显的提升。

然而，由于缺乏对其他智能体信息的考虑，智能体通常会在非平稳环境中运行。为了解决分散学习带来的效率低下问题，乐观和延迟优化会给更好的 Q 值分配更大的更新权重，从而缓解非平稳性的问题。循环神经网络等技术也可以缓解这些问题。

例如，使用 DQN 更新策略时，直接使用经验重放缓冲区中的数据会加剧非平稳性，此时可以采用重要性抽样给不同的样本分配不同的权重来解决这个问题。

$$\mathcal{L}(\theta_i)=\frac{\pi^{t_c}{}_i(a_{-i}|o_{-i})}{\pi^{t_i}{}_i(a_{-i}|o_{-i})}[(y_i^Q-\boldsymbol{Q}_i(o_i,a_i;\theta_i))^2] \tag{1}$$

其中，θ_i 是智能体 i 的 Q 网络参数，t_c 是当前时间，t_i 是样本采集时间，y_i^Q 是时间差目标。

3.2 集中训练集中执行

如图 2(b)所示，在 CTCE 框架中，每个智能体只负责将观测值 o_i 发送给中央控制器，并不包含策略网络；中央控制器负责所有智能体学习集中式联合策略以及决策并相应执行，因此中央控制器中包括 n 个智能体的策略网络 $\pi(a_i|o_1,\cdots,o_n;\theta_i)$。

因此，在 CTCE 框架中，MAS 的训练可以使用所有的 SARL 算法来进行。然而，算法的复杂性会随着状态和动作的维度呈指数增长[31]，虽然这个问题可以通过策略或价值分解来解决，但是 CTCE 难以评估智能体之间的相互影响。

此外，CTCE 框架因为存在可扩展性差、鲁棒性低、通信成本高、缺乏灵活性等缺点，在实际应用中并不常用。尽管 CTCE 在实际应用中通常会受到一定限制，但在某些模拟环境或小型系统中，当所有智能体都紧密耦合且通信成本较低时，CTCE 不失为一个可行的选择。CTCE 可以将整个系统视为单个智能体，以此来充分利用所有智能体的观察结果[32]。因此，在需要完全通信和共享观察的场景下，如在团队运动（如足球、篮球）类型的场景中，智能体能够观察到全局信息，并且在协作中，智能体之间需要共享其观察信息，此时使用 CTCE 方法可以充分利用所有智能体的观察信息，将整个系统视为一个单一的智能体进行决策[33]。

3.3 集中训练分散执行

CTDE 框架[34]中，在训练阶段，各智能体能够利用所有智能体的全局信息或共享信息，从而更准确地优化智能体的策略：$\pi_i:\Omega_i\to\Delta(A_i)$。由于每个智能体都能观察到其他智能体的行为和环境的变化，因此这种方式有助于解决环境的非平稳性问题，提高训练效率。环境的非平稳性指环境的统计特性会随时间的变化而变化，这要求智能体能够适应环境的动态变化。

而在执行阶段，由于每个智能体包含自身的策略网络 $\pi_i(a_i|o_i;\theta_i)$，如图 2(c)所示，因此智能体能够完全自主地根据其本地观测信息做出决策，而不需要与其他智能体进行实时通信，这使得 CTDE 框架在实际应用中具有较高的鲁棒性、灵活性和可扩展性。

CTDE 框架因为结合了集中训练和分散执行的优点，所以在 MAS 中得到了广泛的应用，特别是在某些复杂场景中表现出色[35]。

虽然 CTDE 框架具有众多优点，但是该范式的高效协作仍然是协作多智能体系统中的一个挑战。在 CTDE 框架下，智能体对环境是部分可观测的，导致无法充分利用全局信息进行集中训练，限制了智能体搜索并收敛到全局最优联合策略[36]。为解决这个问题，Zhou 等[37]基于 CTDE 提出了一种新颖的多智能体强化学习框架——集中建议分散剪枝(CADP)。该方法结合集中训练分散执行的优点，以及显式通信和模型剪枝技术，尝试对 CTDE 进行完全集中的训练，以促进训练期间的明确智能体合作，同时仍然确保执行策略的独立性。在《星际争霸 II》微观管理挑战赛和 Google Research Football 基准测试中进行的实验表明，配备 CADP 框架的各种 MARL 方法产生的结果优于其他最先进的方法。

此外，在 CTDE 中，一个常见的挑战是如何在训练时有效地分配奖励[38]。如果没有明确的机制告诉智能体它们的行动是如何影响团队整体性能的，那么智能体可能会学习到错误的协作策略，导致它们之间的行为倾向不一致，从而产生分歧。简而言之，即 MAS 中的奖励是驱动行为倾向的最重要的指导信号。这种分歧就是源于 CTDE 在奖励分配过程中缺乏足够的团队共识指导信号。Zhang 等[39]提出了内在行动倾向一致性方法来解决这个问题，该方法通过将行动模型获得的内在奖励集成到奖励附加 CTDE(RACTDE)框架中。在 SMAC 和 GRF 基准测试中对具有挑战性的任务进行的实验，展示了该方法的性能改进。

在人机协作智能体训练过程中，考虑到人机之间的交互特点、任务需求、环境复杂性以及安全性等多个方面，CTDE 由于在训练阶段可以利用全局信息进行集中训练，且在执行阶段独立执行，有助于智能体在考虑到人类的行为模式和任务目标，学习到与人类协同工作时的最优策略的同时，提高 MAS 的灵活性和鲁棒性。因此，CTDE 适用于在高度人机协作前提下自主智能体训练的场景，并且可以从同构智能体扩展到异构智能体环境。

但在 CTDE 框架下，智能体之间的通信和计算成本过高，训练过程相对复杂，策略泛化能力较差，因此在解决环境非平稳性问题的前提下，使用 DTDE 作为人机协作中智能体的训练范式可能是一个更好的选择。

3 种训练范式的对比如表 2 所列。

表 2 MARL 训练范式的对比

Table 2 Comparison of MARL training paradigm

训练范式	优点	缺点	适用场景
分散训练分散执行(DTDE)	通信成本低，独立性强	环境非平稳性，协作性差，策略冲突	异质的多智能体环境
集中训练集中执行(CTCE)	全局优化，策略一致性	通信成本高，可扩展性差，中心化依赖	需要完全通信和共享观察的场景
集中训练分散执行(CTDE)	结合了 DTDE 和 CTCE 的优点：高效协作，可扩展性好	训练与执行的差异，实现复杂	非平稳性的环境

4 人机协作中智能体的训练方法

在人机协作中，智能体必须与人类伙伴适当地交流知识、意图和目标，同时正确理解和解释伙伴的知识、意图和目标，并将其个体行动与人类的行动相协调，才能成功实现任务目标[40]。此外，智能体与人类协作需要快速适应人类的个人优劣势和偏好，而大多数标准的多智能体强化学习技术产生的智能体过于适合其训练伙伴，导致泛化能力差，即不能很好地推广到不同水平的人类。因此在智能体的训练过程中，泛化能力是很重要的评估标准[41]。

除此之外，智能体在训练时是否采用人类数据也是一个很重要的问题。设计与人类伙伴互动协作的智能体时，需要考虑人类伙伴的偏好和期望[42]。因此在前期的一些研究中，研究人员会收集大量的人类数据来训练智能体。虽然这种方法可以提高智能体的泛化能力，但是收集并使用大量人类数据会使得智能体的训练任务繁重、成本昂贵，且大量使用人类数据样本会导致人类隐私泄露。对此，基于无人类数据，即零样本(zero-shot)的方法被提出，其具有较好的泛化能力。尽管一些研究表明，在某些环境中，智能体可能无需使用人类数据进行训练就可以与真实的人类协作，但在人类行为的微妙特征对任务产生严重影响的情况下，不可能在没有人类数据的情况下生成有效的协作策略。因此，在考虑训练成本问题和人类偏好的情况下，可以使用少量人类数据，即少样本(few-shot)。

本文根据人类数据样本使用情况，将人机协作中智能体的训练方法分为 3 类：有人类数据样本、无人类数据样本、少人类数据样本。

有人类数据样本的智能体训练方法可以帮助智能体根据人类的反馈进行训练，以更好地适应人类的决策流程。该类方法主要适用于人类主导型的人机协作。无人类数据样本的智能体训练方法可以使智能体学会在没有人类指导的情况下做出有效决策，从而具备较强的自主决策能力。该类方法主要聚焦于智能体自主型的人机协作[8]。少人类数据样本的智能体训练方法可以帮助智能体在有限的人类交互数据下，快速适应并参与到协作中。该类方法主要适用于人机平等型的人机协作。

4.1 有人类数据样本

在人机协作中，智能体需要适应人类伙伴的偏好，并且要对不同的人类伙伴有较好的泛化能力，这就需要在智能体的训练过程中使用大量的人类数据样本。人类样本提供了真实世界中的各种场景、情景和事物，这些数据能够帮助智能体更好地理解和分析各种情况，从而提高预测的准确性和可靠性。例如，在自动驾驶中，智能体需要具有较高的人类相似性（对其他交通参与者的礼貌以及对预测不确定性的信心）以保证能够安全有效地与其他实体（人类驾驶员、行人或其他自动驾驶智能体）互动，比如在十字路口谨慎驾驶[43]。此外，人类数据样本的多样性和复杂性可以使智能体更好地应对各种复杂和多变的情况。通过训练，智能体能够学习到更多的边界条件和异常情况，从而增强其鲁棒性和稳定性。

前期的一些研究中，解决两人零和博弈问题的一种有效的方法是让一个智能体与一组其他的智能体（通常是自身的过去版本）一起玩来训练。但是这种类似自我博弈或群体博弈的方法并不能很好地扩展到协作型智能体的训练中，即使在某些研究中表现良好，这些优势也可能是来自智能体或 AI 系统的自身能力，而不是与人类的协作能力。

Carroll 等[4]指出，允许共同训练的智能体可能会收敛到不透明的协作策略，当这样的智能体与人类协作时，它将执行不透明的策略，当人类不发挥作用时，该策略可能会导致严重的失败。因此，他们假设，在真正的协作场景中，经过训练后可以与其他智能体良好协作的智能体在与人类协作时表现会更差；随后他们进一步假设，将人类数据样本纳入训练过程将带来显著的改进。在 Overcooked 游戏环境中的实验验证了上述假设，即使使用简单的行为克隆模型，也能大幅改善代理的协作表现。

行为克隆（Behavior Cloning，BC）是机器学习的一种技术，特别适用于模仿学习中。在行为克隆中，智能体通过观察人类专家或高质量策略的演示来学习行为策略。具体来说，智能体从专家演示中提取状态到动作的映射，并尝试复制这种映射，以在类似的环境中执行任务。BC 已成功应用于许多领域[5]，在人机协作中可以用于生成在智能体模拟训练期间充当人类伙伴的模型，这种方法称为行为克隆博弈（Behavior Cloning Play，BCP）。

BCP 作为一种最典型的使用人类数据样本训练智能体的方法，首先使用 BC 技术，通过收集人类数据来训练一个虚拟人类模型，然后利用该模型来训练“具有人类意识”的智能体，如图 3 所示。但 BCP 存在一个问题，即它假设使用的人类数据集中包含了所有可能出现的情况，但是如果在执行过程中出现未见过的情况，则训练出的智能体可能会表现不佳，且不能很好地泛化到不同的人类伙伴。这也是过度依赖人类数据样本训练智能体最大的问题之一。此外，由于 BCP 或类似使用人类数据样本的方法依赖于大量的标注数据和相似环境，这类方法缺乏解释性和鲁棒性。解决这些问题最好的办法是增加样本数量，但这又带来了另一个问题，即训练的成本会大大增加，且大量使用人类数据样本也会导致人类隐私泄露等问题。

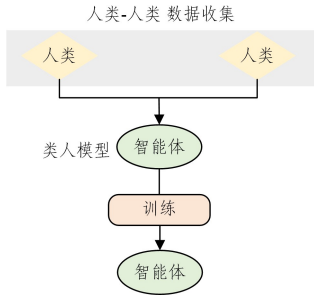


图 3 行为克隆训练

Fig. 3 Behavior cloning play

4.2 无人类数据样本

在前期的研究中，人机协作中智能体的训练都使用大量的人类数据样本来使智能体更好地泛化到人类伙伴，但是这种方法由于涉及收集大量的人类数据，因此训练的步骤繁重且昂贵，并且很容易泄露人类的隐私。基于这些问题，一些研究在智能体的训练期间不使用人类数据样本，如引入在竞争性环境中使用并取得成功的 SP 和 PP 等方法[44]。这些方法最初被应用于两人零和博弈，在围棋[45]、雷神之锤[46]、Dota 和星际争霸等游戏中取得了很大的成功，给研究人员留下了深刻的印象。最初在人机协作智能体的训练中，SP 或 PP 这类方法是有效提高协作性的一种手段，智能体通过自我协作或群体协作进行训练，不断增强协作能力。

SP 方法如图 4（a）所示。智能体在训练过程中从自身副本进行的重复游戏中学习，这种方法训练出的智能体并不能很好地推广到新的人类合作伙伴[47]，原因是经过 SP 训练的智能体只学会与自己协作，因此与行为不同的新合作伙伴建立的合作关系比较脆弱且策略的选择很单一[48]。PP 方法如图 4（b）所示。训练一组智能体后，所有智能体都相互交互[49]。虽然 PP 可以生成在竞争性团队游戏中与人类协作的智能体，但它仍然无法在纯粹的共同收益环境中（其中所有智能体都朝着共同的目标努力并获得相同的奖励）为新的人类伙伴生成强大的协作伙伴[46]。在共同收益环境中，PP 会鼓励智能体以相同的方式进行训练，因此会减少策略多样性并生成与 SP 没有太大区别的智能体[50]。综上所述，直接将传统竞争性环境中训练零和博弈智能体的方法引入到协作环境中时，智能体的泛化能力是一个很大的挑战。

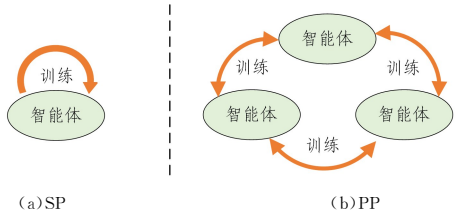


图 4 自我博弈和群体博弈

Fig. 4 Self-play and population-play

零样本协调（Zero-shot Coordination，ZSC）是近年来针对协作多智能体任务提出的概念，其主要目标是构建可以与从未见过的新伙伴（如人类）进行协调的 AI 智能体[51]，即智能体能够在没有事先共同训练或明确约定的情况下，通过一次性的交互就实现与新伙伴有效的协作[52]。这种能力对于

实现自然、灵活且高效的人机协作系统至关重要[53]。

在 ZSC 框架中，两个玩家各自独立地训练一个联合策略(Joint-policy)，并在训练后通过交叉博弈(Cross-play，XP)评估它们的策略。假设 π_1 和 π_2 是两名玩家独立产生的联合策略，那么它们的 ZSC 性能就是在从每个联合策略中匹配单个组件时获得的平均 XP 回报。这一目标可以表示为：

$$J_{XP}(\pi_1,\pi_2)=\frac{1}{2}(J(\pi_1^1,\pi_2^2)+J(\pi_2^1,\pi_2^1)) \tag{2}$$

在基于 ZSC 的人机协作智能体训练方法的研究中，研究者提出并解决了如下问题。

4.2.1 独立决策问题

在完全合作问题中，由于 SP 智能体在训练过程中控制自己的轨迹分布，因此每种策略通常仅在此精确分布上表现良好。这导致在与另一个智能体一起游戏时，它们很可能会陷入训练过程中未遇到的情况。综上，SP 的一个主要缺点是它未能充分模拟现实中独立个体在解决协作任务时所需的独立决策过程。这一局限性促进了 ZSC 问题的研究[54]。在 ZSC 问题中，智能体必须独立地为协作任务生成策略，且这些策略能够与训练期间未见过的合作伙伴兼容。

Hu 等[55]将 ZSC 问题具体化为寻找一种通用学习算法，这种算法能够使独立训练的智能体在测试时成功地进行协调。独立训练的智能体在这里被视为人类在解决协调任务时所需的具有独立决策能力的智能体。在现实生活中，人们往往需要基于各自的知识、经验和策略做出决策，而不是通过共享模型或策略来进行。因此，独立训练能够模拟这种个体间的独立性和自主性。他们假设每个玩家只同意使用一种学习算法，但不共享环境中观察、动作和状态的标签。以杠杆协调游戏为例，如图 5 所示，两个智能体可以从 10 个不同的杠杆中选择。如果两个智能体选择了相同的杠杆，它们将获得相应的奖励。其中有一个杠杆的奖励是 0.9，其余的奖励是 1。如果智能体选择了不同的杠杆，则奖励为 0。在这种情况下，SP 算法将学习一个联合策略，即两个智能体都选择奖励为 1 的杠杆。但在没有杠杆标签的情况下，智能体无法区分具有同等奖励的杠杆，这样的策略是无法协调的。

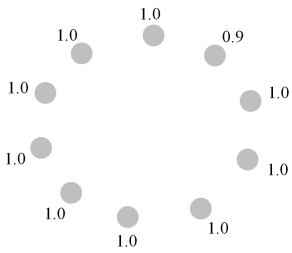


图 5 杠杆协调游戏中的奖励

Fig. 5 Rewards in the lever coordination game

为解决上述问题，Hu 等[55]提出了其他博弈(Other-play，OP)算法。该算法的核心思想是利用问题中已知的对称性来增强传统的 SP 方法，以寻找更鲁棒的策略。例如，在杠杆协调游戏中，导致奖励为 1 的动作都是对称的，因此随机排列的策略将以相同的概率选择它们。相比之下，导致收益为 0.9 的杠杆没有对称的对应物，因此可以始终选择该杠杆。实验中，在与人类玩家进行 Hanabi 游戏时，OP 算法的性能优于 SP 算法，这说明了 ZSC 对人机协作的益处。然而，他们并没有将“无标签”假设正式化，而是依赖于其证明中的直观概念。

为正式化定义“无标签”，Treutlein 等[54]提出了无标签协调(Label-free Coordination，LFC)，并定义了无标签协调游戏。研究表明，OP 不是 LFC 问题的最佳解决方案，它可能会在独立训练中学习不同的、可能不兼容的最优策略，导致不能有效选择不同智能体相容的最优策略。因此他们提出了 OP 算法的扩展，即打破平局的 OP(OP with Tie-breaking)算法，并证明它是 LFC 问题的最佳解决方案。

Lupu 等[56]在解决独立策略问题时，提出了对一组智能体训练一个共同的最佳响应，并通过调控保持智能体群体的多样化的方法。为此，他们引入了一种用于生成多样化强化学习策略的可微分目标函数——轨迹多样性(Trajectory Diversity，TrajeDi)，并将 TrajeDi 推导为策略之间 Jensen-Shannon 散度的泛化，然后在一个简单的矩阵游戏中通过实验验证了其有效性，发现它能够找到唯一的 ZSC 最优解。最终实验结果表明，使用 TrajeDi 目标函数可以找到多样化的解决方案，特别是在两个树状环境中，智能体能够学习到不同的路径以达到高奖励状态，从而在 ZSC 框架中表现更好。

4.2.2 过度拟合问题

在协作任务中，与单个队友一起进行大量训练可能导致对特定队友行为风格的“过度拟合”问题。传统的多智能体强化学习方法(SP 或 PP)，往往导致智能体过度拟合其训练伙伴，难以泛化到新的人类合作者[44]。使用 BCP 训练智能体虽然可以解决这个问题，但又会产生需要收集大量人类数据、成本高且繁琐的问题。

如何解决智能体的过拟合问题，提高其泛化能力，成为一个研究重点。在智能体必须泛化到新的人类伙伴的情况下，实现临时、零样本协调尤其重要。许多成功的方法都采用了人类模型，要么显式构建[57]，要么隐式学习[58]。相比之下，竞争领域的最新研究表明，在没有人类数据的情况下，通过 SP 方法，使用无模型强化学习可以达到人类水平[59]。基于此，Strouse 等[6]提出了“没有人类数据的无模型强化学习能否生成可以与新的人类伙伴协作的代理?”这个问题，并且基于虚拟自我博弈(Fictitious Self-play，FSP)提出了一种名为“虚拟协同博弈”(Fictitious Co-Play，FCP)的方法。如图 6 所示，该方法通过训练智能体作为 SP 智能体及其训练过程中各个检查点($t=1$，$t=2\cdots$)的最佳响应，来产生多样化的训练伙伴，从而提高智能体与新的人类或智能体伙伴协作的效果。

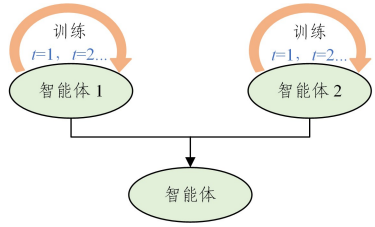


图 6 虚拟协同博弈

Fig. 6 Fictitious co-play

与新的伙伴协作的一个重要挑战是处理对称性[55]。例如，面对面站立的两个智能体 A 和 B 可以通过 A 向左移动而

B向右移动来相互通过,反之亦可。这两种都是有效的解决方案,但如果人类明显偏爱其中一种方式,那么一个好的智能体伙伴就会在这些约定之间进行适应性切换。第二个重要挑战是处理技能水平的差异,即好的智能体伙伴应该能够协助不同水平的合作伙伴。

Strouse 等[6]认为鲁棒的智能体合作者的关键在于与多样化的训练伙伴进行训练,即智能体泛化能力可以通过训练更大、更多样化的群体来提升。在 FCP 方法中,他们训练了 N 个 SP 代理作为一个多样化的伙伴池,仅改变它们神经网络的随机初始化种子。在训练过程中,为了使该池能够代表不同的技能水平,定期保存代表该时刻策略的智能体作为"检查点"。最终检查点代表了一个完全训练的"熟练"伙伴,而早期检查点则代表了"不太熟练"的伙伴。然后,训练一个 FCP 智能体伙伴,使其对经过全面训练的智能体及其过去的检查点做出最佳反应。训练过程中,伙伴的参数是固定的,FCP 必须学会适应伙伴,而不是期望伙伴来适应它。该方法中,不同的检查点模拟了不同的技能水平,而不同的随机种子则以不同的方式打破了对称性。后续在 Overcooked 环境中的实验表明,FCP 方法在与不同类型的合作伙伴(包括人类和智能体)合作时的表现显著优于其他基线方法(SP,PP,BCP),在各种测试情况下都取得了更高的分数。在与人类合作时,FCP 方法表现最好,不仅在客观表现上优于其他方法,在主观偏好上也受到人类参与者的青睐。

但是,由于未能解决合作环境中智能体之间基于特定共同知识的训练导致的问题,即智能体在训练过程中倾向于利用这些共同知识,从而在面对未见过的伙伴时无法有效协作,因此在完全合作的多智能体马尔可夫决策过程中,广泛使用的如伙伴抽样(Partner Sampling)和基于人群的训练(Population-based Training,PBT)等方法引入的多样性是不可靠的。为此,Charakorn 等[60]提出,生成预训练合作伙伴是一种简单且有效的方法,可以在完全合作的环境中引入多样性,进一步的测试中也证实了预训练智能体在实际的临时团队中的表现更优秀。Sarker 等[61]还提出,传统的多智能体强化学习技术(如 SP)往往会导致智能体收敛到任意且不具多样性的约定,使得智能体的泛化能力受到限制,为此他们提出通过最大化自我博弈奖励,同时最小化与先前发现的约定的奖励,来生成多样化的约定,从而提高智能体在面对不同伙伴时的泛化能力。

另一方面,现有的采用 SP 方法的研究基本都假设任务是同构的,无法很好地捕捉智能体和人类伙伴之间的合作行为,因此对于异构 ZSC 任务效率低下。然而,许多现实世界的任务是异构的,即智能体和合作伙伴具有不同的优势或技能。例如,机械臂可以举起人类无法举起的重物,而人类可以进行比机械臂更灵活的操作。异构设置具有更好的适用性,并且在必要时可以恢复同质策略[62]。Xue 等[63]在 Strouse 等[6]的基础上提出了"能否在没有人类数据的情况下生成能够有效地与新的和异质的人类协调的智能体?"的问题,即异构 ZSC 问题。此外,还提出了一种基于协同进化的通用方法,即通过协同进化进行多智能体零样本协调(MAZE)。该方法通过配对、更新和选择这 3 个子过程,将智能体暴露给不同的合作伙伴,来共同进化两个智能体和伙伴群体,使智能体能够在 ZSC 中更好地协调。在每一代中,来自两个群体的智能体和伙伴首先配对,以在环境中进行协调,然后通过优化奖励和多样性的加权和来进行更新。为了进一步增加伙伴的多样性,MAZE 维护了一个存档,用于存储过去生成的各种伙伴,并从中选择一些过去的伙伴来训练。

这些方法在解决因与单个队友一起训练而导致的对特定队友行为风格的潜在过度拟合问题上取得了很大的成功。

4.2.3 分布偏移问题

当通过 SP 训练获得的智能体与未遇到的伙伴(如人类)配对时,它们可能会受到分布偏移的严重影响,导致其性能下降。分布偏移是指智能体在训练时所使用的数据分布与实际应用或测试时遇到的数据分布之间的差异,这种差异会显著影响智能体的性能。

为了缓解这种分布变化,Zhao 等[64]提出了基于最大熵群体的训练(Maximum Entropy Population-based Training,MEP)。在 MEP 中,群体中的智能体通过推导出的群体熵奖励进行训练,以促进智能体之间的成对多样性和智能体自身的个体多样性,并通过优先采样与这个多样化的群体中的智能体配对,来训练一个共同的最佳智能体,如图 7 所示。这种优先级的设置会根据训练进度动态调整。在强化学习的环境中,熵通常用来衡量策略的随机性或多样性。一个高熵的策略意味着在给定状态下,智能体选择不同动作的概率分布较为均匀,即动作选择的不确定性较高;一个低熵的策略则意味着某些动作被选择的概率远高于其他动作,动作选择的不确定性较低。

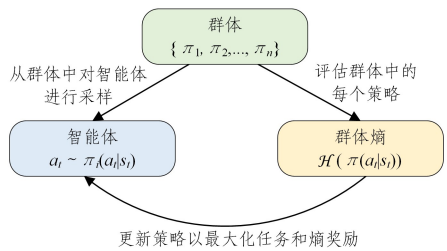


图 7 基于最大熵群体的训练

Fig. 7 Maximum entropy population-based training

标准的 RL 旨在最大化奖励的期望总和 $E_\tau[\sum_t R(s_t, a_t)]$,在学习开始时,几乎所有动作都具有相等的概率。而经过一些训练后,一些动作在累积更多奖励的方向上具有更高的概率。随后,在训练过程中,策略的熵会随时间减小[65]。而最大熵 RL 则在标准 RL 的目标上增加了策略的期望熵[66],这激励了智能体选择非主导动作。最大熵强化学习的目标定义为:

$$J(\pi)=\sum_t E_{(s_t,a_t)\sim\pi}[R(s_t,a_t)+\alpha\mathcal{H}(\pi(\cdot|s_t))] \tag{3}$$

其中,参数 α 调节熵奖励与奖励之间的相对重要性。

为了鼓励群体中个体智能体策略的多样性和可探索性,对个体策略使用最大熵目标。该方法将个体多样性和成对多样性的组合定义为群体多样性(PD),并推导出一个安全且计算高效的替代目标群体熵(PE),它是具有线性运行时复杂度的原始 PD 目标的下界。与最大熵强化学习训练类似,群体

中的每个智能体都会获得奖励，以最大化群体熵。通过这个多样化的群体，在基于协作难度的优先级方案下，将一个最佳响应智能体与从该群体中抽样的智能体配对并进行训练。如此，这个新训练的智能体会遇到一组不同的策略，并且可以具有更好的泛化能力。

1)群体多样性

基于最大熵强化学习，我们希望群体中的策略具备探索性和多样性。首先，通过使用最大熵的奖励，可以鼓励每个智能体的策略本身具有探索性和多样性。其次，通过将群体中每个策略对的 KL 散度作为目标的一部分，来鼓励群体中的策略互补且互不相同。

因此，将群体多样性定义为每个智能体策略的平均熵和群体中每个智能体对之间的平均 KL 散度的组合：

$$PD(\{\pi^{(1)},\pi^{(2)},\cdots,\pi^{(n)}\},s_t):=\frac{1}{n}\sum_{i=1}^{n}\mathcal{H}(\pi^{(i)}(\cdot\mid s_t))+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}D_{\mathrm{KL}}(\pi^{(i)}(\cdot\mid s_t),\pi^{(j)}(\cdot\mid s_t)) \tag{4}$$

其中，KL 散度(D_{KL})和熵($\mathcal{H}$)的定义如下：

$$D_{\mathrm{KL}}(\pi^{(i)}(\cdot\mid s_t),\pi^{(j)}(\cdot\mid s_t))=\sum_{a\in A}\pi^{(i)}(a_t\mid s_t)\log\frac{\pi^{(i)}(a_t\mid s_t)}{\pi^{(j)}(a_t\mid s_t)} \tag{5}$$

$$\mathcal{H}(\pi^{(i)}(\cdot\mid s_t))=-\sum_{a\in A}\pi^{(i)}(a_t\mid s_t)\log\pi^{(i)}(a_t\mid s_t) \tag{6}$$

但是，由于 KL 散度是无界的，因此，优化这个目标作为奖励函数的一部分可能会导致收敛问题。

2)群体熵

为了提高 PD 目标的稳定性和运行时的复杂性，推导一个有界且有效的优化智能体目标 PE，PE 被定义为群体策略的平均熵：

$$PE(\{\pi^{(1)},\pi^{(2)},\cdots,\pi^{(n)}\},s_t):=\mathcal{H}(\bar{\pi}(\cdot\mid s_t)) \tag{7}$$

其中：

$$\bar{\pi}(a_t\mid s_t):=\frac{1}{n}\sum_{i=1}^{n}\pi^{(i)}(a_t\mid s_t) \tag{8}$$

将 PE 目标作为 PD 目标的下界。

3)基于最大熵的群体训练

借助 PE 目标，可以训练一批能够良好合作且策略互不相同的智能体。因此，与最大熵强化学习中的目标类似，将 MEP 的训练目标定义如下：

$$J(\bar{\pi})=\sum_t E_{(s_t,a_t)\sim\pi}[R(s_t,a_t)+\alpha\mathcal{H}(\bar{\pi}(\cdot\mid s_t))] \tag{9}$$

其中，$\bar{\pi}$ 是群体的平均策略，α 确定了群体熵相对于任务奖励的相对权重。

在获得最大熵群体后，再利用这组多样化的智能体来训练一个协作智能体，它可以轻松地与任何协作伙伴协作。

4.2.4 零样本协调能力评估

智能体的 ZSC 能力通常通过几个评估伙伴(包括人类和代理)进行评估，并以平均回报进行报告。当前针对 ZSC 能力的评估方法仍需在构建多样化的评估伙伴和全面衡量 ZSC 能力方面进行改进[67]。具体来说，广泛使用的人类代理智能体(Human Proxy Agents)与人类并不相似[68]。

Wang 等[69]为创建一种可靠、全面且高效的 ZSC 能力评估方法，正式定义了理想的“多样性完整”评估伙伴，并提出了最佳响应(Best Response，BR)多样性，即伙伴策略的最佳响应所构成的群体多样性，用来近似理想的评估伙伴。他们提出了一种评估流程，包括“多样性完整”评估合作伙伴构建和多维指标，即最佳响应接近度(Best Response Proximity，BR-Prox)指标[70]。BR-Prox 将 ZSC 能力量化为与每个评估伙伴的近似 BR 之间的性能相似度，从而展示了 ZSC 的泛化能力和改进潜力。BR-Prox 的正式定义为：

$$BR\text{-}Prox(\pi,\{\pi_{\omega^i}\}_{i\in p}):=Aggr_{i\in p}(U(\pi,\pi_{\omega^i})/U(\hat{BR}(\pi_{\omega^i}),\pi_{\omega^i})) \tag{10}$$

其中，$Aggr$ 表示评估合作伙伴之间的聚合器，如最常见的“均值”和“中位数”聚合器。在这里，使用四分位均值作为聚合器，即中间 50% 数据的均值，这种聚合器在统计上是可靠的[71]。

在后续实验中，Wang 等[69]使用该评估流程重新评估了在 Overcooked 环境下的一些强大的 ZSC 方法，发现在一些常用的布局中，现有方法的性能无法区分，并且现有 ZSC 方法在生成足够多样且高性能的训练伙伴方面也存在不足。这些结果说明了该评估流程的有效性，其为高效评估 ZSC 方法提供了一个有益补充。

综上所述，各类问题的本质目标都是在不适用人类数据样本的情况下，提高协作智能体的泛化能力，这也是人工智能面临的长期挑战[72]。

除此之外，ZSC 问题的研究还包括训练范式设计、等变网络设计[73]、基于策略相似性评估的协调增强[74]、ZSC 问题的一般场景、基于集成技术的 ZSC 改进、人类价值偏好研究[68]等。

4.3 少人类数据样本

人机协作的本质仍然需要智能体将人类参与者视为合作实体。尽管上述无人类数据样本训练协作智能体的研究表明，在某些环境中，智能体可能无需对人类数据进行训练就可与真实人类进行协作[6]，但人类行为的微妙特征会对任务产生严重的影响，在这种情况下，若不使用人类数据样本训练智能体，则会导致智能体不能生成有效的协作策略。

传统上，通过 BCP 从人类演示中学习可以产生高性能策略，前提是算法能够访问大量的高质量数据，这些数据涵盖了智能体在操作时最可能遇到的场景。然而，在真实场景中，专家数据是有限的，且使用人类数据样本的训练方法又引出了成本高、过程繁杂和人类隐私泄露等问题，这就使得研究者不得不考虑，如何使得智能体的训练方法既能保证高效、泛化，又能减少成本。为此，很多研究在基于少样本(Few-shot，FS)[8]的前提下，使智能体在人机协作中产生高性能策略。Fosong 等[75]基于 FS 提出了少样本团队合作(Few-shot Teamwork，FST)的问题，即将训练有素、擅长完成某一任务的智能体与来自不同任务领域的技能熟练的智能体结合起来，共同学习以适应一个未见但相关的任务。他们将 FST 问题视为解决两个独立问题的过程：1)减少训练一个智能体团队，以完成复杂任务所需的经验；2)实现与陌生队友的合作，以完成新任务。

在 FST 问题中，训练过程分为两个阶段：源阶段和调整

阶段。在源阶段,每个子团队被训练来完成一个相对简单的源任务。在调整阶段,这些子团队被组合成完成目标任务的整个团队,并进行有限的训练以适应新任务。Fosong 等[75]提出了两种不同的 FST 问题定义框架:1)基于课程学习的方法,旨在最小化训练目标任务所需的总经验;2)基于临时团队协作的方法,旨在最小化调整阶段所需的训练步数。

将 FST 引入人机协作时,可以使用少量不同的人类数据样本训练多个智能体,再将它们整合起来,共同学习来适应不同的人类伙伴,实现低成本训练单个智能体,且训练出来的智能体泛化能力强。

Ding 等[76]提出了协调方案探测(Coordination Scheme Probing,CSP)方法以解决在 ZS 环境中的过拟合问题。CSP 不进行 ZSC,而是应用一个方案探测模块,以利用预先收集的有限情景数据来预先捕获未知队友的协调方案。为了实现泛化,CSP 以端到端的方式学习了一个元策略,其中包含多个遵循不同协调方案的子策略,并自动重用该策略来与未见过的队友进行协调。当与人类协作时,CSP 方法可以主动收集少量人类数据来预学习其人类伙伴的策略,从而实现更好的协作。因此,CSP 方法具有泛化到不同人类伙伴的能力。

先前的研究主要集中在单一任务或多任务场景下的协调能力提升,忽略了任务以持续方式出现的情况。为此,Yuan 等[77]提出通过渐进式任务情景化实现多智能体持续协调(Multi-Agent Continual Coordination via Progressive Task Contextualization,MACPro)方法来解决这一问题。该方法的关键在于获得一个分解策略,该策略使用共享的特征提取层,但具有分离的独立任务头,每个任务头专门处理特定类别的任务。任务头可以根据学习任务情景化进行渐进式扩展。此外,考虑到主流的协作式 MARL 中流行的 CTDE 训练范式,在集中式训练过程中,利用每个智能体的局部信息,通过策略蒸馏来近似策略头选择过程,智能体可以选择最优的策略头,以分散的方式与其他队友协调。

这些利用少样本技术来应对多模态场景的研究已经取得了初步成功,并表现出较好的性能,且泛化能力强。

Islam 等[78]表示,在智能系统中利用人类的专长和经验与 AI 相结合既高效又有益,并且提出人类可以向 RL 智能体提供有用的建议,从而使它们能够在多智能体环境中改进学习。在后续的研究中,他们也证明了智能体向人类学习是有效的,并且人类与智能体的合作在复杂的模拟环境中优于人类控制或完全自主的 AI 智能体。

Waytowich 等[79]指出,自动课程学习(Automatic Curriculum Learning)作为一种新机制,拥有根据智能体的当前能力调整当前要解决的任务的难度来加速深度强化学习的技术。然而,为足够复杂的任务设计适当的课程可能具有挑战性。因此,他们利用人类演示作为在训练期间指导智能体探索的一种方式。这项工作的目标是训练深度强化学习智能体,这些智能体可以指挥多个异构执行者,其中起始位置和任务的总体难度由单个人类演示生成的自动课程控制。他们的研究结果表明,通过自动课程学习训练的智能体在《星际争霸 II》中模拟的指挥和控制任务中的表现优于最先进的深度强化学习基线,并且与人类专家的表现相匹配。

Nekoei 等[7]的研究发现,当前竞争的 ZSC 算法在面对不同的学习方法时需要相当数量的样本来适应新的队友。因此,他们提出了一种少样本协作(Few-shot Collaborative,FSC)方法,并验证了该算法在 Hanabi 上的有效性。

对于要训练的智能体,另一种方法是通过先验偏差直接对具有人类行为风格的队友进行编码[18],或者将基于先验偏差的手动编码人类行为与使用人类交互数据的优化智能体相结合[80]。

表 3 人机协作智能体的训练方法

Table 3 Training methods of agents in human-agent collaboration

分类		训练方法/算法	方法/算法简介	局限性
有人类数据样本		BCP	首先收集人类数据,使用 BC 技术训练一个虚拟人类模型,然后利用该模型来训练“具有人类意识”的智能体	泛化能力差,缺乏解释性和鲁棒性,训练成本高,易泄露人类隐私
无人类数据样本	独立决策问题	OP	通过利用问题中已知的对称性来增强传统的 SP 方法,以寻找更鲁棒的策略	不能有效选择不同智能体相容的最优策略
		OP with tie-breaking	OP 算法的扩展,即打破平局的 OP 算法	在某些情况下,可能不存在合理的 tie-breaking 函数,而是需要学习一个完全不同的策略
		TrajeDi	对一组智能体训练一个共同的 BR,并通过调控保持智能体群体的多样化	在更复杂的环境中的表现还有待进一步探索
	过度拟合问题	FCP	通过训练智能体作为 SP 智能体及其训练过程中各个检查点的最佳响应,来产生多样化的训练伙伴,从而提高智能体的泛化能力	研究中假设任务是同构的,无法很好地捕捉智能体和人类伙伴之间的合作行为
		MAZE	通过配对、更新和选择这 3 个子过程将智能体暴露给不同的合作伙伴,以共同进化两个智能体和伙伴群体,使智能体能够在 ZSC 中更好地协调	MAZE 只使用了简单的随机配对策略,可以考虑通过课程学习等方法进一步提高配对的效率
	分布偏移问题	MEP	群体中的智能体通过推导出的群体熵奖励进行训练,以促进智能体之间的成对多样性和智能体自身的个体多样性,并通过优先采样与该多样化的群体中的智能体配对,来训练一个共同的最佳智能体	可能不适用于所有类型的游戏或任务,特别是具有非常特殊动态的游戏或任务
少人类数据样本		CSP	应用一个方案探测模块,以利用预先收集的有限情景数据来预先捕获未知队友的协调方案	需要额外的环境交互和训练步骤
		MACPro	该方法的一个分解策略,该策略使用共享的特征提取层,但具有分离的独立任务头,每个任务头专门处理特定类别的任务	任务情景化学习和任务头选择机制有待优化,以提高 MACPro 在大规模任务场景下的可扩展性

5 人机协作的应用与发展

人机协作作为当前科技发展的重要方向，正在引领工业、医疗、教育等多个领域的深刻变革。智能化、人性化和安全化将是人机协作未来的发展趋势。随着人工智能技术的不断进步，智能体将具备更强大的自主学习和决策能力，从而更好地与人类协作。

在人机协作智能体的构建中，AI 智能体框架作为构建自主感知与决策智能体的基础架构，起着至关重要的作用。当前的 AI 智能体框架简化创建、部署与管理，削减复杂度，促进自动化与智能化任务执行，优化团队协作，是智能系统构建的关键。

大模型的出现为 AI 智能体提供了“聪明的大脑”，并重新定义了 AI 智能体。当前，由大模型驱动的 AI 智能体架构是比较常见的 AI 智能体落地架构，包含规划(Planning)、记忆(Memory)、工具(Tools)、执行(Action)四大要素，如图 8 所示。AI 智能体能够充分发挥大模型的能力，尤其是在解决复杂问题方面，相较于传统 AI 应用架构实现了跃迁级提升。

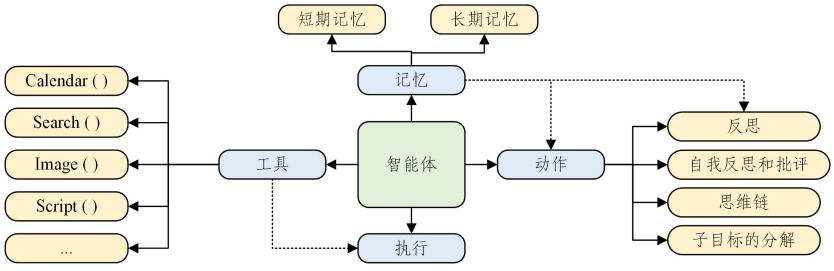


图 8 由大模型驱动的 AI 智能体架构

Fig. 8 AI agent frameworks driven by large model

LangChain，LangGraph，CrewAI 与 AutoGen 作为前沿 AI 框架，在人机协作智能体的构建中发挥着重要作用。LangChain 通过链式编程，将 AI 功能串联，简化了复杂任务的处理，同时增强了 AI 系统的自然语言交互能力，促进了人机之间的有效沟通与合作[81]。LangGraph 作为 LangChain 的扩展，其图形化结构和多智能体能力使开发者能构建出复杂且动态的应用程序，支持多角色、多步骤的协作流程，极大地提升了人机协作的流畅性和灵活性。CrewAI 专注于智能体间的无缝协作，通过精细的角色分配和任务管理，使 AI 智能体能够共同解决复杂任务[82]。在人机协作中，CrewAI 的动态协作机制显著提高了任务执行效率，为复杂场景提供了智能化支持。AutoGen 通过自动化代码和文档生成，简化了编程流程，减少了人为错误，使开发者能更专注于创意与策略。在人机协作中，AutoGen 的实时监控与调整功能确保了任务高效准确地执行，同时其多智能体系统与人机交互深度融合，提升了协作的整体效能。具体如表 4 所列。

表 4 4 种主流 AI 智能体框架的对比

Table 4 Comparison of four mainstream AI agent frameworks

框架	特点	优势	适用场景
LangChain	基于 LLM 的应用	多功能性，外部集成	通用 AI 智能体开发
LangGraph	有状态的多角色系统	复杂的工作流程，智能体协调	交互式自适应 AI 应用
CrewAI	角色扮演 AI 智能体	解决合作问题，高度灵活兼容	模拟复杂的组织任务
AutoGen	多智能体对话系统	稳健性，模块化，易用性	先进的对话式 AI 和任务自动化

在人机协作智能体的训练中，从最初的使用大量人类数据样本，到后续研究不使用人类样本，以减少初期如 BCP 等训练方法带来的成本高昂等问题，再到后来发现不使用人类数据样本(如 ZSC 方法)的训练方法会带来泛化能力低等问题，而引出使用少样本结合智能体自主学习来训练智能体的方法。大量的研究都表明：在人机协作领域，想要智能体能够更好地适应不同的人类合作伙伴，且减少训练繁杂程度或成本，一定要减少人类数据样本的使用。因此后续的研究可以结合使用人类数据样本和 ZSC 的方法来实现智能体与人类的更好协作，如将 BCP 方法和 FCP 方法进行结合研究。随着 SP，PP，FCP[6,46] 和 CSP[76] 等一系列研究取得成功，智能体训练环境的问题也逐渐成为重点。

自计算机诞生之初，游戏就一直是研究智能体如何进行复杂决策的重要测试环境。在人机协作方面，尽管近年的研究取得了一些进展，但该领域仍面临诸多挑战，最典型的是缺乏便捷有效的训练及测试环境。当前，大多数人机协作研究的测试主要依赖于 Overcooked 等[4]少数游戏或平台。然而，这些平台往往智能体数量有限，场景单一，难以全面反映真实世界中人机协作的复杂性和多样性。因此，开发更多样化、更贴近实际应用的测试环境成为当务之急[8]。

当下人机协作领域的研究大多在虚拟游戏环境中进行，这导致现实环境中存在的一些问题并不能很好地在智能体训练中得到泛化。因此，随着虚拟环境中的人机协作研究获得更大的成功，将人机协作的研究引入现实环境是一个重要的挑战。

在未来的研究中，不仅仅要考虑智能体训练方法的优化，一个好的、全面的测试环境也是重点的研究问题，且测试环境的提升是下一步研究的前提和基础。

人机协作技术正处于快速发展阶段，并将继续在多个领域与人类实现更密切的合作。随着技术的不断进步，人机协作将带来更多创新、机会和挑战，推动工作方式的变革，开创更加繁荣和可持续的未来。

结束语 近年来，人机协作不断发展，以 ZSC 为基础的研究提出了众多无人类数据样本的研究方法，这些方法不仅

在实际实验中取得了巨大的成就,还为后续的研究提供了基础。与传统的使用人类数据样本的方法行为克隆博弈相比,ZSC方法实现了自然、灵活且高效的人机协作系统。随着ZSC的进一步发展,FSC被提出,以解决ZSC不能很好地适应人类行为偏好的问题。此外,如何更好地实现少样本协调是一个亟待解决的问题,后续可以考虑结合行为克隆博弈(BCP)和ZSC的优势来实现FSC,以解决当前ZSC存在的问题。未来,人机协作将会在各个领域得到广泛应用,人机协作智能体与人类合作伙伴的合作能力也会在后续的研究中进一步深化。

参 考 文 献

[1] GAO Y,LIU F,WANG L. Towards Effective and Interpretable Human-Agent Collaboration in MOBA Games: A Communication Perspective[C]// Proceedings of the International Conference on Learning Representations. Kigali:ICLR,2023:1-26.

[2] DORRI A,KANHERE S S,JURDAK R. Multi-Agent Systems: A Survey[J]. IEEE Access,2018,6:28573-28593.

[3] CHENS,WANG Y,SONG Z,et al. WHALES: A Multi-agent Scheduling Dataset for Enhanced Cooperation in Autonomous Driving[J]. arXiv:2411.13340,2024.

[4] CARROLL M,SHAH R,HO M K,et al. On the Utility of Learning about Humans for Human-AI Coordination[C]// Proceedings of the Neural Information Processing Systems. Vancouver:NeurIPS,2019:5175-5186.

[5] BAIN M,SAMMUT C. A Framework for Behavioural Cloning [C]// Proceedings of the Machine Intelligence 15. Oxford:Oxford University Press,2000:103-129.

[6] STROUSE D,MCKEE K R,BOTVINICK M,et al. Collaborating with Humans without Human Data[C]// Proceedings of the Neural Information Processing Systems. Virtual Event: NeurIPS,2021:14502-14515.

[7] NEKOEI H,ZHAO X T,RAJENDRAN J,et al. Towards Few-shot Coordination:Revisiting Ad-hoc Teamplay Challenge In the Game of Hanabi[C]// Proceedings of the Conference on Lifelong Learning Agents. Montral:CoLLAs,2023:861-877.

[8] YUAN L,ZHANG Z,LI L,et al. A Survey of Progress on Cooperative Multi-agent Reinforcement Learning in Open Environment[J]. arXiv:2312.01058,2023.

[9] KAELBLING L P,LITTMAN M L,MOORE A W. Reinforcement Learning: A Survey[J]. Journal of Artificial Intelligence Research,1996,4:237-285.

[10] LI Y X. Deep Reinforcement Learning An Overview[J]. arXiv: 1810.06339,2018.

[11] WONG A,BÄCK T,KONONOVA A V,et al. Deep Multi-agent Reinforcement Learning:Challenges and Directions[J]. Artificial Intelligence Review,2022,56(6):5023-5056.

[12] OROOJLOOY A,HAJINEZHAD D. A Review of Cooperative Multi-agent Deep Reinforcement Learning[J]. Applied Intelligence,2022,53(11):13677-13722.

[13] GRONAUER S,DIEPOLD K. Multi-agent Deep Reinforcement Learning:A Survey[J]. Artificial Intelligence Review,2022,55: 895-943.

[14] EKER B,OZKUCUR E,MERICLI C,et al. A Finite Horizon DEC-POMDP Approach to Multi-robot Task Learning[C]// Proceedings of the 2011 5th International Conference on Application of Information and Communication Technologies(AICT). Baku:IEEE Xplore,2011:1-5.

[15] YANG Y,WANG J. An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective[J]. arXiv: 2011.00583,2020.

[16] BERKEL N V,SKOV M B,KJELDSKOV J. Human-ai Interaction:Intermittent,Continuous and Proactive[J]. Interactions, 2021,28(6):67-71.

[17] ONNASCH L,ROESLER E. A Taxonomy to Structure and Analyze Human-Robot Interaction[J]. International Journal of Social Robotics,2021,13(1):833-849.

[18] PUIG X,SHU T,LI S,et al. Watch-And-Help:A Challenge for Social Perception and Human-AI Collaboration[C]// Proceedings of the International Conference on Learning Representations. Vienna:ICLR,2021:1-23.

[19] AJOUDANI A,ZANCHETTIN A M,IVALDI S,et al. Progress and Prospects of the Human-robot Collaboration[J]. Autonomous Robots,2017,42(5):957-975.

[20] GOODRICH M A,SCHULTZ A C. Human-Robot Interaction: A Survey[J]. Foundations and Trends in Human-Computer Interaction,2007,1(3):203-275.

[21] MICHALOS G,KARAGIANNIS P,DIMITROPOULOS N, et al. The 21st century industrial robot:When tools become collaborators[M]. Berlin:Springer International Publishing,2021, 17-29.

[22] YANG G,ZHOU H Y,WANG B C. Digital Twin-driven Smart Human-machine Collaboration: Theory, Enabling Technologies and Applications[J]. Journal of Mechanical Engineering,2022, 58(18):279-291.

[23] ABRAMSON J,AHUJA A,BRUSSEE A,et al. Imitating Interactive Intelligence[J]. arXiv:2012.05672,2020.

[24] MANGAL U,MOGHA S,MALIK S. Data-Driven Decision Making: Maximizing Insights Through Business Intelligence, Artificial Intelligence and Big Data Analytics[C]// Proceedings of the 2024 International Conference on Advances in Computing Research on Science Engineering and Technology. Indore:IEEE Xplore,2024:1-7.

[25] HADDADIN S,CROFT E. Physical Human-Robot Interaction [M]. Berlin: Springer International Publishing, 2016: 1835-1874.

[26] LUCK M,MARK D. A Conceptual Framework for Agent Definition and Development[J]. Computer Journal,2001,44:1-20.

[27] HENTOUT A,AOUACHE M,MAOUDJ A,et al. Human-robot interaction in industrial collaborative robotics: a literature review of the decade 2008-2017[J]. Advanced Robotics,2019, 33(15/16):764-799.

[28] KOLBEINSSON A,LAGERSTEDT E,LINDBLOM J. Foundation for a classification of collaboration levels for human-robot cooperation in manufacturing[J]. Production & Manufacturing Research,2019,7(1):448-471.

[29] HAESEVOETS T,CREMER D,DIERCKX K,et al. Human-

Machine Collaboration in Managerial Decision Making[J]. Computers in Human Behavior, 2021, 119: 106730.
[30] WU X, CHANDRA R, GUAN T, et al. iPLAN: Intent-Aware Planning in Heterogeneous Traffic via Distributed Multi-Agent Reinforcement Learning[J]. arXiv: 2306.06236, 2023.
[31] FOSTER D J, FOSTER D P, GOLOWICH N, et al. On the Complexity of Multi-Agent Decision Making: From Learning in Games to Partial Monitoring[C]// Proceedings of the Annual Conference Computational Learning Theory. Bangalore: COLT, 2023: 2678-2792.
[32] CHEN Y, YANG W, ZHANG T, et al. Commander-soldiers reinforcement learning for cooperative multi-agent systems[C]// Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN). IEEE, 2022: 1-7.
[33] LIU C, LIU G. JointPPO: Diving Deeper into the Effectiveness of PPO in Multi-Agent Reinforcement Learning[J]. arXiv: 2404.11831, 2024.
[34] OLIEHOEK F A, SPAAN M T J, VLASSIS N. Optimal and Approximate Q-value Functions for Decentralized POMDPs[J]. Journal of Artificial Intelligence Research, 2008, 32: 289-353.
[35] LYU X, BAISERO A, XIAO Y, et al. On Centralized Critics in Multi-Agent Reinforcement Learning[J]. Journal of Artificial Intelligence Research, 2023, 77: 295-354.
[36] WANG J, YE D, LU Z. More Centralized Training, Still Decentralized Execution: Multi-Agent Conditional Policy Factorization [C]// Proceedings of the International Conference on Learning Representations. Kigali: ICLR, 2023: 1-18.
[37] ZHOU Y, LIU S, QING Y, et al. Is Centralized Training with Decentralized Execution Framework Centralized Enough for MARL? [J]. arXiv: 2305.17352, 2023.
[38] MATIGNON L, LAURENT G J, LE FORT-PIAT N. Independent Reinforcement Learners in Cooperative Markov Games: A Survey Regarding Coordination Problems[J]. The Knowledge Engineering Review, 2012, 27(1): 1-31.
[39] ZHANG J, ZHANG Y, ZHANG X S, et al. Intrinsic Action Tendency Consistency for Cooperative Multi-Agent Reinforcement Learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2024: 17600-17608.
[40] MUTLU B, TERRELL A, HUANG C M. Coordination Mechanisms in Human-Robot Collaboration[C]// Proceedings of the HRI 2013 Workshop on Collaborative. 2013.
[41] MCKEE K R, LEIBO J Z, BEATTIE C, et al. Quantifying the Effects of Environment and Population Diversity in Multi-agent Reinforcement Learning[J]. Autonomous Agents and Multi-Agent Systems, 2022, 36(1): 1-16.
[42] DAFOE A, HUGHES E, BACHRACH Y, et al. Open Problems in Cooperative AI[J]. arXiv: 2012.08630, 2020.
[43] WANG L, SUN L, TOMIZUKA M, et al. Socially-Compatible Behavior Design of Autonomous Vehicles With Verification on Real Human Data[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 3421-3428.
[44] WANG X, TIAN Z, WAN Z, et al. Order Matters: Agent-by-agent Policy Optimization[J]. arXiv: 2302.06205, 2023.
[45] SILVER D, HUBERT T, SCHRITTWIESER J, et al. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm[J]. arXiv: 1712.01815, 2017.
[46] JADERBERG M, CZARNECKI W M, DUNNING I, et al. Human-level Performance in 3D Multiplayer Games with Population-based Reinforcement Learning [J]. Science, 2019, 364(6443): 859-865.
[47] LOWE R, GUPTA A, FOERSTER J, et al. On the Interaction Between Supervision and Self-play in Emergent Communication [J]. arXiv: 2002.01093, 2020.
[48] BULLARD K, KIELA D, PINEAU J, et al. Quasi-Equivalence Discovery for Zero-Shot Emergent Communicatio[J]. arXiv: 2103.08067, 2021.
[49] LANCTOT M, ZAMBALDI V, GRUSLYS A, et al. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning[C]// Proceedings of the Neural Information Processing Systems. Long Beach: NeurIPS, 2017: 4190-4203.
[50] GARNELO M, CZARNECKI W M, LIU S, et al. Pick Your Battles: Interaction Graphs as Population-Level Objectives for Strategic Diversity[C]// Proceedings of the Autonomous Agents and Multiagent Systems. Virtual Event: AAMAS, 2021: 1501-1503.
[51] KLEIMAN-WEINER M, LITTMAN M L, TENENBAUM J B, et al. Coordinate to Cooperate or Compete: Abstract Goals and Joint Intentions in Social Interaction [EB/OL]. [2016-08-10]. https://mindmodeling.org/cogsci2016/papers/0295/index.html.
[52] SHUM M, KLEIMAN-WEINER M, LITTMAN M L, et al. Theory of Minds: Understanding Behavior in Groups through Inverse Planning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 6163-6170.
[53] LERER A, PEYSAKHOVICH A. Maintaining Cooperation in Complex Social Dilemmas Using Deep Reinforcement Learning [J]. arXiv: 1707.01068, 2017.
[54] TREUTLEIN J, DENNIS M, OESTERHELD C, et al. A New Formalism, Method and Open Issues for Zero-Shot Coordination [C]// Proceedings of the International Conference on Machine Learning. Virtual Event: ICML, 2021: 10413-10423.
[55] HU H, LERER A, PEYSAKHOVICH A, et al. "Other-Play" for Zero-Shot Coordination[C]// Proceedings of the International Conference on Machine Learning. Virtual Event: ICML, 2020: 4399-4410.
[56] LUPU A, HU H, FOERSTER J. Trajectory Diversity for Zero-Shot Coordination[J]. Adaptive Agents and Multi-Agent Systems, 2021, 139: 7204-7213.
[57] CHOUDHURY R, SWAMY G, HADFIELD-MENELL D, et al. On the Utility of Model Learning in HRI[C]// Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE: Daegu, 2019: 317-325.
[58] SADIGH D, LANDOLFI N, SASTRY S S, et al. Planning for Cars that Coordinate with People: Leveraging Effects on Human Actions for Planning and Active Information Gathering over Human Internal State[J]. Autonomous Robots, 2018, 42: 1405-1426.
[59] BROWN N, SANDHOLM T. Superhuman AI for Multiplayer Poker[J]. Science, 2019, 365(6456): 885-890.
[60] CHARAKORN R, MANOONPONG P, DILOKTHANAKUL

N. Investigating Partner Diversification Methods in Cooperative Multi-agent Deep Reinforcement Learning[M]. Bangkok: Springer International Publishing,2020:395-402.

[61] SARKAR B,SHIH A,SADIGH D. Diverse Conventions for Human-AI Collaboration[C]// Proceedings of the Neural Information Processing Systems. New Orleans:NeurIPS,2023:1-25.

[62] KUBA J,FENG X,DING S,et al. Heterogeneous-Agent Mirror Learning:A Continuum of Solutions to Cooperative MARL[J]. arXiv:2208.01682,2022.

[63] XUE K,WANG Y,YUAN L,et al. Heterogeneous Multi-agent Zero-Shot Coordination by Coevolution[J]. arXiv:2208.04957,2022.

[64] ZHAO R,SONG J,YUAN Y,et al. Maximum Entropy Population-Based Training for Zero-Shot Human-AI Coordination[J]. Proceedings of the AAAI Conference on Artificial Intelligence,2023,37(5):6145-6153.

[65] MNIH V,ADRIÈ PUIGDOMÈNECH B,MIRZA M,et al. Asynchronous Methods for Deep Reinforcement Learning[C]// Proceedings of the International Conference on Machine Learning. New York:NeurIPS,2016:1928-1937.

[66] HAARNOJA T,ZHOU A,ABBEEL P,et al. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor[C]// Proceedings of the International Conference on Machine Learning. Stockholm: NeurIPS, 2018: 1856-1865.

[67] CHARAKORN R,MANOONPONG P,DILOKTHANAKUL N. Generating Diverse Cooperative Agents by Learning Incompatible Policies[C]// Proceedings of the International Conference on Learning Representations. Kigali:ICLR,2023:1-15.

[68] YU C,CHAO J X,LIU W L et al. Learning Zero-Shot Cooperation with Humans, Assuming Humans Are Biased[J]. arXiv: 2302.01605,2023.

[69] WANG X,ZHANG S,ZHANG W,et al. Quantifying Zero-shot Coordination Capability with Behavior Preferring Partners[J]. arXiv:2310.05208,2023.

[70] KIRK R,ZHANG A,GREFENSTETTE E,et al. A Survey of Zero-shot Generalisation in Deep Reinforcement Learning[J]. Journal of Artificial Intelligence Research,2023,76:201-264.

[71] AGARWAL R,SCHWARZER M,CASTRO P S,et al. Deep Reinforcement Learning at the Edge of the Statistical Precipice [C]// Proceedings of the Neural Information Processing Systems. Virtual Event:NeurIPS,2021:29304-29320.

[72] KNOTT P,CARROLL M,DEVLIN S,et al. Evaluating the Robustness of Collaborative Agents[C]// Proceedings of the Adaptive Agents and Multi-Agent Systems. UK: AAMAS, 2021: 1560-1562.

[73] MUGLICH D,WITT C S D,VAN DER POL E,et al. Equivariant Networks for Zero-Shot Coordination[C]// Proceedings of the Neural Information Processing Systems. New Orleans: NeurIPS,2022:6410-6423.

[74] LI Y,ZHANG S,SUN J,et al. Cooperative Open-ended Learning Framework for Zero-shot Coordination[J]. International Conference on Machine Learning,2023,202:20470-20484.

[75] FOSONG E,RAHMAN A,CARLUCHO I,et al. Few-Shot Teamwork[J]. arXiv:2207.09300,2022.

[76] DING H,JIA C,GUAN C. Coordination Scheme Probing for Generalizable Multi-agent Reinforcement Learning[C]// Proceedings of the ICLR 2023 Conference Blind Submission. 2023.

[77] YUAN L,LI L,ZHANG Z,et al. Multi-agent Continual Coordination via Progressive Task Contextualization[J]. arXiv:2305.13937,2023.

[78] ISLAM S,DAS S,GOTTIPATI S K,et al. Human-AI Collaboration in Real-World Complex Environment with Reinforcement Learning[J]. arXiv:2312.15160,2023.

[79] WAYTOWICH N R,HARE J,GOECKS V G,et al. Learning to Guide Multiple Heterogeneous Actors From A Single Human Demonstration via Automatic Curriculum Learning in StarCraft II[J]. arXiv:2205.05784,2022.

[80] SHIH A,SAWHNEY A,KONDIC J,et al. On the Critical Role of Conventions in Adaptive Human-AI Collaboration[J]. arXiv: 2104.02871,2021.

[81] BHATT A,NANDAN V. Med-Bot:An AI-Powered Assistant to Provide Accurate and Reliable Medical Information[J]. arXiv:2411.09648,2024.

[82] DUAN Z,WANG J. Exploration of LLM Multi-Agent Application Implementation Based on LangGraph+CrewAI[J]. arXiv: 2411.18241,2024.

HUANG Weiye, born in 1999, postgraduate. His main research interest is multi-agent reinforcement learning.

CHEN Xiliang, born in 1985, Ph.D, associate professor. His main research interests include command information system engineering and deep reinforcement learning.

(责任编辑:柯颖)