多模态环境中的多智能体强化学习：预训练大模型视角

预印本，v0.2，2023 年 3 月

作者：温颖，万梓煜，张劭，张伟楠

摘要：现实世界本质是一个动态开放的多模态多智能体环境，生物与环境的不断共同进化造就了其无尽的复杂度。因此，智能体需要在现实的开放式环境中不断迁移泛化已有能力，与其他智能体或人类不断交互，持续适应未知场景，以实现更高水平的智能。在本文中，作者尝试从开放式的多模态多智能体环境中两个关键要素出发：1）基于预训练大模型，进行开放式的多模态信息进行理解、落实与利用；2）智能体与其他决策主体之间（包括其他智能体和人类）的竞争与合作；总结已有多智能体强化学习的研究成果，并展望未来方向。

关键词：强化学习，开放式环境，多模态，预训练大模型，人智协同

1. 引言

自上世纪 50 年代人工智能（Artificial Intelligence，AI）的概念被首次提出以来，人们常常畅想一个通用人工智能（Artificial General Intelligence，AGI）的存在，试图最大化由人给定的期望效用，可以在现实世界中完成特定任务。面向这个目标，研究者在过去几十年经历了从逻辑推理、专家系统、统计机器学习到深度学习的多轮探索。到 2011 年之后，深度学习技术在算力和数据的支撑下，分别在语音识别、计算机视觉和自然语言处理方面取得巨大进步。2016 年，DeepMind（英国人工智能公司）提出 AlphaGo，借助深度强化学习技术，首次击败了世界围棋冠军，进一步推动了人工智能研究的发展。这些研究通过给智能体定义明确的学习目标并提供相应的训练数据集或模拟器，在特定任务上取得了良好的效果，但同时也存在着数据样本利用效率低与策略泛化性能差等问题。**然而，现实场景中的任务越来越复杂，需要智能体具备更高的认知与自主学习能力，才能应对不断变化的任务场景，难以假设智能体拥有固定的目标，亟需更一般化的学习设定与理论，推动人工智能研究的进一步发展。**

为实现更通用的人工智能，智能体需要具备良好的泛化性能和与其他决策主体进行交互决策的能力。由于现实环境与任务存在较大的复杂性，同时一直处于动态变化的过程中，我们难以对每一个任务设定具体的学习目标，导致人工智能只能应用于小范围特定任务上。因此，通过预训练，在海量数据上学习到足够丰富的先验知识，可以为人工智能提供更好的决策基础，并能够快速适应新任务。此外，在现实场景中，

智能体需要与其他不同目标或偏好的决策主体进行竞争与合作，导致学习目标不断变化，智能体难以适应。这意味，自主产生任务并设定目标，以应对不同的任务或对手。同时，智能体的目标与行动需要能与人类价值对齐，以保证不偏离人类的目标，不对人类造成损害。而当前单一任务中，人类的价值都是预先设定的先验知识，然而在现实环境中人类的价值是不断变化的，这导致传统的人工智能很难与人类在现实环境中达成价值对齐。

开放式多模态多智能体环境是接近真实世界的智能体学习环境和设置，为探索智能体泛化和交互决策能力提供了良好的试验场，是研究更通用人工智能技术的重要方向之一。这个环境与现实世界有着一致的多模态感知信息和大量开放式任务，智能体可以在其中与其他决策主体交互。在开放式多模态多智能体环境中，智能体需要处理和融合来自不同模态的信息，例如文本、图像和语音等，以获取更丰富的信息和先验知识。并在与其他决策主体的交互中学习适应不同的任务和目标。智能体还需要具备自主探索和学习的能力，以适应不同的任务和目标，并与其他决策主体进行协作或竞争。在这种开放式多模态多智能体环境下，智能体的学习目标更加不确定和多样化，需要通过新的学习方法和范式来解决样本效率、策略泛化和竞争协作等问题。

预训练大模型和强化学习技术的发展为实现通用人工智能提供了新的思路。以 ChatGPT 为代表的自然语言模型及一些列图像-文本双模态的预训练模型，通过在海量无标注数据上进行训练，充分利用大数据、大模型和大算力的优势，从而突破小任务的学习范式，形成新的预训练-提示/微调的学习范式。强化学习则针对样本效率的研究，使用基于模型的方法、模仿学习与离线强化学习等方法，在特定任务上较实现了针对离线数据利用效率的大幅度提升。在多智能体学习方面，基于种群的方法在竞争博弈场景中获得良好的经验效果。在这个基础上，我们可以想象，**在预训练好的具备大量世界先验知识的预训练大模型上，智能体可以从混沌中寻找到规律，结合多智能体强化学习，与虚拟或现实环境进行交互决策，实现开放式多模态环境中进行自主探索和学习，从规律中创造秩序，是一个提升智能体泛化能力可行路径。**然而，现阶段开放式多模态环境中的多智能体强化学习研究才刚刚起步，如何使智能体在现实开放式的多模态场景进行自主探索与学习，与其他决策主体进行交互，是帮助智能体从游戏走向更广泛场景，从虚拟走向现实，最终实现更加通用人工智能的关键步骤。

为此，本文围绕智能体学习如何在开放式多模态环境中进行交互学习的问题展开，尝试对已有工作进行梳理，总结研究进展，并展望未来研究方向。面向这个目标，我们认为不能仅局限在现有强化学习方法中，而是需要为开放式环境中的智能体学习研发新的学习方法和范式，包括结合多智能体学习、自动课程、博弈论和自监督等方法，同时依靠预训练大模型、自然语言理解与计算机视觉等领域前沿进展的推动，最终实现智能体在开放式环境中能够持续自主学习的目标。如图 1 所示，本文将从智能体在开放式环境中的学习目标与可感知信息量两个维度讨论，尝试对开放式环境中的智能体学习研究工作进行阶段性总结：

一、多模态信息感知、理解、认知与生成角度。真实的开放式环境，场景动态变

化，通常包含不同任务或者与任务不直接相关的多模态信息。我们将着重讨论如何利用自然语言处理、计算机视觉和预训练大模型的前沿成果，处理和融合多模态信息，提高策略泛化性与样本效率。

二、智能体学习目标构建与价值对齐角度。在单智能体场景下，学习目标通常是最大化长期回报。即使是多任务学习的目标，也是可以被明确定义为最大化各项任务上的累计或平均回报。但在智能体与其他决策主体（包括智能体或人类）存在交互的情况下，学习目标很难在学习开始之前就被明确定义。根据智能体与其他决策主体的交互博弈类型分类，从智能体竞争、智能体合作到人智协同（Human-AI Collaboration），智能体学习目标的不确定性越来越大。如何在多智能体场景下，定义智能体明确的学习目标或让智能体能够自主发现目标，是实现智能体快速自适应的关键。因此，我们强调智能体需要能探索学习目标，对齐人类价值，以解决智能体自主学习与竞争协作能力，更好地帮助人类。

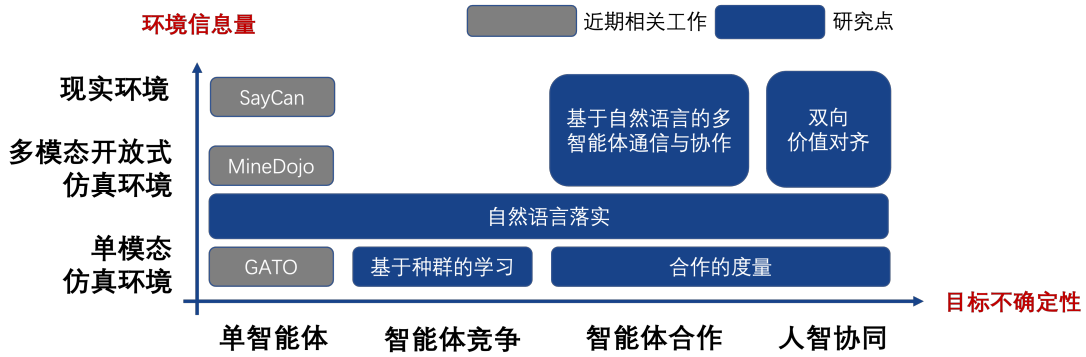

图 1 开放式多模态环境中多智能体学习的研究维度

2. 开放式多模态环境中的多智能体学习

现实世界作为一个开放式环境，本质上是一个由各种任务、目标及生物组成，处于不断动态变化的过程中，并持续产生无数新的任务和目标，同时这些新的任务和目标又不断促进生物（包括人在内）的学习与进化，以最终实现生存的目的。因此，现实世界具有多样化的特征，包括多模态的输入信息，需要智能体能够感知和理解不同的信息类型，并与环境本身及环境中其他的决策主体进行交互。现在的智能体还无法达到人类智能的水平，是因为其缺乏在开放式多模态环境中的交互学习能力。在现实世界中，智能体需要能在动态变化的环境、任务和其他自适应主体不断交互过程中保持鲁棒性，并且快速迁移已有知识，适应未知挑战，才能更好地服务于人类需求。

开放式多模态环境中的多智能体学习是指在开放式环境中，通过多个智能体（包括人类）之间的相互作用和协作，使用多模态的输入信息（例如图像、语音、文本等）来实现目标任务的学习过程，与现实世界有着较为接近的设定。开放式环境中，任务和对手是不断动态变化的，需要智能体具备对理解的信息进行抽象的能力，能够进行认知和泛化，对智能体的学习能力提出了更高的要求。在这种学习环境中，智能体需

要根据环境的变化和任务的要求，自主探索不同任务目标与偏好进行学习，以最大程度地实现能力的复用泛化。因此，智能体需要具备对信息进行抽象和理解的能力。语言是一种对现实世界的抽象描述方式。因此，可以将动态变化环境中的信息抽象到语言上，从而实现对任务和目标的认知和泛化。这将有助于智能体在不同的任务和场景下快速适应和学习，并具备更好的泛化性和自适应性。

当然，仅仅对语言层次的抽象是不够的。多模态环境环境意味着智能体需要处理来自不同传感器和数据源的信息，如图像、语音、文本等。因此，多模态智能体的学习需要解决如何融合多模态信息的问题，以及如何实现策略的泛化等挑战。此外，现实世界中不仅仅需要考虑与环境的交互，还需要考虑到多个智能体之间的相互作用，进行竞争与协作。在多智能体学习中，智能体需要与其他决策主体合作和竞争，以实现任务目标和优化总体性能。这种合作和竞争过程是不断变化和动态的，需要智能体具备快速适应和学习的能力，以及对任务和目标的认知和理解。其中，最困难的挑战是我们不知道“问题”在哪里。一旦“问题”或者目标被明确了，就能有一个清晰的方向。因此一个重要的探索的方向是如何实现学习算法的自动“进化”，用人工智能创造出更好的人工智能，在开放式环境中获得更好的泛化性和自适应性。同时，多智能体学习也需要对齐人类的价值观，以确保智能体的学习过程是符合人类需求和利益的。因此，多智能体学习需要解决如何协调多智能体之间的决策、如何实现策略的泛化等挑战，以及如何确保学习过程符合人类价值观的问题。

3. 为什么我们需要预训练大模型？

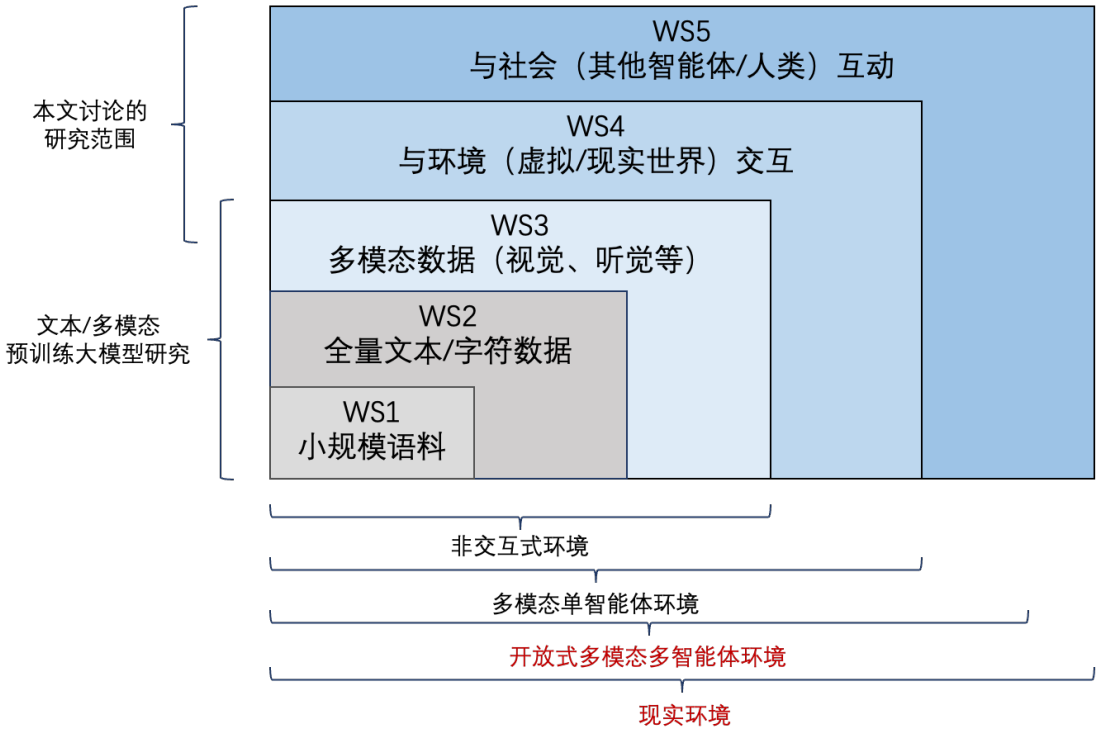

图 2 自然语言学习的世界范围概念示意图

随着预训练大模型的发展，尤其是自然语言大模型取得了突破性成果，在训练过程中对大量真实世界语料的学习以及抽象，让其具备了对现实世界不同模态信息进行表达的能力，这使得开放式多模态环境中的多智能体强化学习研究变得可行。人们开始将自然语言预训练大模型和强化学习相结合，这种组合可以为强化学习提供更好的决策基础。接下来，我们讨论为什么需要在强化学习的基础上结合预训练大模型。

强化学习现阶段的预训练方式无法实现泛化。目前，强化学习的泛化能力和样本效率都存在问题。这是因为以往的强化学习任务领域特定，抽象程度不够，轨迹信息量不高，导致无法建立通用的预训练模型。在这种情况下，强化学习模型的泛化性能和样本效率都受到了限制。如果要在更广泛的应用场景中使用强化学习，就必须提高其泛化性能和样本效率，这就需要一种更好的预训练基础来提供更好的决策基础。**自然语言预训练大模型为强化学习提供了泛化的可能性与基础。**如图 2 所示，Bisk 等学者提出的“世界范围”(World Scope，WS)概念，阐述了自然语言模型学习的数据源范围。自然语言是一种具有表示性的语言，可以表示宇宙中包含的各种不规则。在 WS1-WS2 范围内，自然语言预训练大模型是在人类对世界高度抽象后的符号化表达数据上训练的模型。进一步，在 WS3 范围，引入自然语言可以理解与描述复杂多样的多模态感知信息，进一步构建多模态的预训练大模型。这些预训练模型可以为强化学习提供更好的决策基础，提高其泛化性能和样本效率。通过将自然语言预训练大模型和多智能体强化学习结合起来，在 WS4-WS5 范围，使得模型与虚拟或物理世界，其他智能体或人进行交互，可以提高强化学习的泛化能力和样本效率，提供更好的决策基础，使其更适用于各种交互决策场景。

强化学习与自然语言大模型的结合能够让智能体认知物理世界，突破语言的边界获得与虚拟或真实世界互动的能力。自然语言大模型由于其对世界的高度抽象以及自然语言理解的能力，具备了从自然语言输入到形式化语言如程序接口的翻译能力，可以将来自真实世界的形式化语言天然转换成可执行的动作/指令。这些指令与动作为强化学习提供了泛化的方向。强化学习借助自然语言大模型来将现实世界与指令/动作对齐，同时利用收集到的人类反馈来提升大模型对世界的感知能力，进而扩展预训练大模型的能力边界。强化学习可以帮助预训练大模型有效利用更多的反馈信息，帮助其更好地理解世界和任务，结合人类反馈来达成对人类价值的对齐。强化学习与自然语言大模型的结合赋予了智能体直接与世界和任务进行交互的能力，智能体可以学习使用工具，执行动作，而自然语言大模型借助强化学习将可以利用获取到的人类反馈进一步提升对世界的理解和认知。这种正向循环可以不断提高预训练大模型的能力，并使其适用于更广泛的应用场景。

4. 预训练大模型驱动的环境感知与动作生成

现有的预训练大模型可以对世界/任务进行符号化抽象，但训练时并没有直接与世界/任务交互过。当自然语言预训练模型与强化学习相结合时，通常需要将环境感知信息转化为预训练模型可接受的输入，并通过预训练模型给出可行的指令，最终指令可

以被策略完整执行并从环境中得到奖励反馈，进一步修正策略与预训练大模型。这其中，利用环境反馈可以进一步纠正预训练大模型对世界/任务的抽象和理解，并利用预训练大模型提高策略性能和泛化性。接下来，我们将从多模态感知信息与预训练大模型的结合，基于预训练大模型的指令生成及基于自然语言的多智能体通信三个角度，展开详细讨论。

4.1 感知融合：多模态信息的自然语言定位/对齐

如同人类通过视觉、听觉和触觉感知和认识世界，开放式环境中智能体可能时刻在感知许多变化的信息，例如物理参数、动作空间、任务目标的自然语言描述、摄像头的视觉信息、传感器的反馈数据等。如何处理与利用这些多模态信息，对开放式环境中的各项任务，例如视觉导航、物体操作和自动驾驶实现可泛化策略学习至关重要。

随着预训练大模型的技术的发展，一系列多模态虚拟环境中智能体自主学习研究逐渐兴起，尤其是在视觉导航和物体操作领域。2022 年 4 月，谷歌机器人（Google Robotics）团队发布了 SayCan，智能体通过使用自然语言预训练模型生成给定任务的子任务指令，然后再根据子任务执行相应动作最终完成任务目标，使得机器人在真实的环境中通过视觉感知执行给定自然语言描述的任务。DeepMind 公司于 2022 年 6 月推出的 GATO，将多步决策任务、多轮对话和图片-文本生成任务统一到一个基于 Transformer 的自回归问题上，并在近 700 个多模态任务上取得了良好表现，初步验证了多模态预训练大模型在智能体学习任务上的潜力。2022 年 7 月，基于我的世界（Minecraft）这一开放世界游戏，英伟达最新发布的 MineDojo 提供了游戏视频，维基百科和论坛网页等多模态数据，同时创建了上千个单智能体任务。MineDojo 基于 CLIP（Contrastive Language-Image Pre-Training）预训练模型在视频-文本对数据集上进行对比学习，训练得到了自然语言目标到环境视觉状态的相关性的零样本（Zero-shot）预测模型 MineCLIP。MineCLIP 可以在开放式环境中根据感知到的视觉状态，评估自然语言所描述任务完成度。同时，任务完成相关度作为奖励信号较一般的强化学习任务奖励可以进一步加速强化学习的训练。从近期工作可以看出，多模态信息通过预训练模型可以在智能体学习中被有效应用，同时自然语言在其中扮演着至关重要的角色，我们将在接下来的章节中展开介绍。

4.2 语言落实：动作抽象与指令生成

不同于自然界中其他生物通过模仿或试错学习技能的方式，人类可以借助自然语言的丰富表达能力对现实世界进行更高级的抽象，产生更复杂的智能行为。因此智能体学习的一个重要的研究点就是自然语言落实（Natural Language Grounding）。自然语言落实是指将自然语言和外部物理世界的丰富的感知连接在一起，从而解决各种多模态问题，同时加强自然语言理解能力。智能体通过自然语言落实，可以将自然语言与感知经验或者行为经验联系起来，如建立语言对于特定物体的描述或是通过简单的语言指令可以完成相应的任务。

因此，随着自然语言预训练大模型的发展，研究者逐渐开始尝试将智能体决策问题中的状态、动作与状态-动作转移空间统一对齐到语言上，从而解决多智能体认知、交流与协作过程中的场景理解，增强智能体策略的合作与泛化能力，并赋予人通过自然语言与智能体直接沟通和协作的可能。近期一些结合自然语言的强化学习研究也证明了这一点，可以通过自然语言对任务目标的描述，以自然语言或半/形式化语言的形式，生成相应步骤的指令。一方面，自然语言可以用作任务目标的描述，常见于指令跟随（Instruction Following）中。在指令跟随任务中，智能体需要根据自然语言的提示或目标完成任务，这类任务在视觉导航（Visual Navigation），游戏，机器人控制中都有广泛的应用。另一方面，自然语言语言可以作为任务的先验知识，一些工作通过从自然语言的先验知识中习得强化学习环境的机制和动态（Dynamics）从而更好的帮助智能体进行策略的探索和训练，比如通过阅读说明手册来完成游戏任务。

4.3 语言涌现：多智能体通信

语言在社会化使用场景中通常用于人与人的交流，其中角色、意图和无数其他变量在某一特定点交织在一起，形成了语言的复杂性。但当前预训练数据与验证场景中往往忽略了这一复杂性。基于语言的通信交互是一种宝贵的信号，但最初的研究由于训练-验证-测试集场景和参考支持的评估而受到限制。为了弥补这一差距，多智能体通信，智能体与人通信，是最终交互式学习的的必要步骤。

多智能体通信是在多智能体间进行去中心化执行时提高合作效率与任务达成表现的常见方式，常被使用在完全合作任务场景中。在多智能体强化学习设定中，智能体之间的通信的研究常用于解决部分可观测问题或基于心智理论（Theory of Mind）促进智能体间的高层次协作。另一个热门的研究方向则是通信涌现（Emergent Communication），旨在通过强化学习来模拟智能体间语言、通信协议的演化过程，并对其性质进行研究。已有的研究通常使用参考游戏（Referential Game），通过“你说我猜”的问答，给出奖励信号，进行通信协议的学习。但在这个基础上学习到的通信协议，在鲁棒性、零样本通信（Zero-shot Communication）以及类语言性质较自然语言都有较大差距。因此，研究者很自然地会思考是否存在一种可能，使得智能体之间的通信能够建立在人类语言模型的先验知识基础上，从而让机器之间的通信自然地拥有人类语言的性质。

受限于以往自然语言模型的建模能力，以往基于自然语言的多智能体通信研究往往效果不佳。但随着预训练的发展，自然语言大模型已经在如文本分类、生成、问答、摘要等领域取得了前所未有的突破。根据大模型比率定理（Scaling Law），越来越多研究者发现当模型参数规模足够大，训练语料足够充足时，预训练语言模型在下游任务上会具有更加优秀的表现， GPT-3（Generative Pre-trained Transformer 3）与其强调的提示学习（Prompt Learning）即是其中具有代表性和开创性的工作之一。因此，在自然语言本身具备对客观世界进行抽象描述能力基础上，再结合自然语言落实，对智能体对行为经验与自然语言的认知对齐，实现智能体间通过自然语言的通信协作，

实现开放式环境中更高效的智能体合作与策略泛化，并通过可解释的交互，进一步促进人智协同。

5. 自动化博弈学习目标构建与人机价值对齐

构建一个能够处理目标不确定性的人工智能系统，需要智能体能自主探索学习目标并能与人类进行协同。已有的方法，通常向人工智能系统事先假定了一个固定已知的目标，所以这些方法可能会需要被重新设计。首先我们考虑学习竞争场景下，由于对抗性对手带来的目标的不确定性，所以需要智能体能够自动发现。其次是合作场景下意图的不确定性。合作是一个竞争更难明确定义的度量。竞争通常拥有明确的胜负的标准，而合作存在不确定性，需要一个更加明确对应。此外，在合作的过程中，人类的意图在开放环境中并非是固定的。人类作为智能，会通过环境与环境中其他智能的行为推断其他智能的意图，从而调整自身的意图与行动。

5.1 竞争博弈：基于种群的学习

在竞争性博弈场景下，研究人员展开了大量开放式学习的研究，以克服智能体学习目标的非传递性（Non-transitive）问题。智能体学习中的非传递性通常指多种策略得到一个或者更多“循环”选择的博弈，这通常会导致智能体无法学习到有效策略。例如在猜拳游戏中，如果剪刀优于布，布优于策略石头，无法推导出剪刀优于石头。因此，竞争性博弈场景下的核心挑战是“问题的问题”（the Problem Problem），即自动化地生成大量多样合适的对手和场景来支撑智能体的学习，可以极大地提高学习效率和效果。

近几年基于种群的学习（Population-based Learning）或者自动课程（Automatic Curriculum）在星际争霸 2、Dota 2 和王者荣耀等多玩家竞争型电子游戏上取得了巨大成功，达到超越人类顶尖玩家的水平。2016 年，DeepMind 公司基于最简单的基于种群算法，自对弈，开发出 AlphaGo ，首次战胜人类最顶级围棋选手。之后，2019 年 1 月，DeepMind 公司提出了基于种群联赛机制的 AlphaStar，在星际争霸 2 全局游戏上与国际最顶尖的选手进行了 11 场对战，并获得了其中 10 场的胜利。与此同时，OpenAI 公司开发出一款基于对手/任务的课程自对弈方法的 OpenAI Five，击败多人在线游戏 Dota 2 上最顶尖的职业玩家。类似地，腾讯公司则在这些研究的基础上加入策略蒸馏对历史经验进行复用，推出王者荣耀绝悟系统，在真实线上测试中击败了99.8%的人类选手。这些成果证明，基于种群的学习不断生成并挑选出合适的对手和任务，可以在较为复杂的多人竞争型电子游戏上取得良好效果。

5.2 合作博弈：合作的度量

当问题从竞争博弈来到合作博弈场景下时，如何衡量智能体之间的合作程度是一个远比竞争场景下衡量竞争优势更为复杂的问题。在非合作博弈的情况下只需要考虑个体的理性，而合作博弈则需要考虑群体的理性。目前，在合作博弈场景下，并没有

一个被广泛接受的合作程度衡量指标或解概念（Solution Concept）。因此，如何确定智能体合作程度，是多智能体合作研究的重要基础。

对于单个智能体的合作能力，我们可以使用固定或动态分布的其他智能体进行评估。在其他智能体分布固定情况下，我们通常采用即兴对弈（Ad-hoc Play）对智能体进行评估，例如 Melting Pot 环境中的配置。研究者可以根据自身研究问题的假设，根据智能体的属性设置合适的分布进行评估。基于动态智能体分布的评估可以是根据不断变化的智能体分布进行评估，更符合开放式环境中的实际情况，例如游戏时每次随机匹配的队友。

由于个体的贡献往往是有限的，很多任务的完成不仅仅取决于单个智能体的个人能力，更依赖于智能体群体的合作能力，因此如何衡量群体的合作水平，对每个智能体的合作能力以及智能体之间的合作程度进行度量，成为了关键所在。在合作环境下，传统自对弈方法（Self-play，SP）往往会导致智能体在训练过程中陷入一些与自身的特殊约定来获取更高的收益，这将最终导致自对弈方法获得的智能体无法与多样化的伙伴合作。基于种群（Population-Based Training，PBT）与虚拟合作博弈（Fictitious Co-Play，FCP）的方法被提出，用于增大训练过程中策略的多样化，以打破这种约定，使得训练获得的策略能够与更多的未知伙伴合作。但这引入了一个全新亟待解决的问题——如何有效评估训练过程中智能体与多样化对手的合作程度，最大化训练的效果，使得智能体能在个人能力与合作能力上都保持持续的提升。这一问题与竞争博弈中学习目标的非传递性是类似的。合作博弈论为解决这一问题提供了一定思路，智能体可以结成联盟，共同合作争取联盟效用最大化，并在联盟内部进行分配的博弈。

5.3 人智协同：双向价值对齐

人工智能的最终目的是希望更好地服务于人类的特定生产生活任务。但现有的智能体学习算法专注于特定任务的效果提升或者智能体与智能体之间的对抗与合作，很少考虑到智能体现实部署过程中与人之间的交互影响。而在现实世界中，人与智能体的关系在大多数时候都是合作的。在开放式环境的合作任务中，人类与智能体的合作效果是决定任务最终表现的关键，而合作不仅是人类适应智能体，智能体也要适应人类。我们也可以将合作效果理解为，人与智能体是否能互相理解对方的意图并进行配合，朝着同一个目标共同前进，推动更好的合作。

人智协同通常可分为三类：1）人类主导，即指由人发出指令，智能体回应人类指令并协助人类完成任务，常见的场景有语音助手以及其他人类可以发出指令的任务；2）智能体主导则相反，由智能体向人类发起请求配合的信号，人类提供信息来完成合作，常见的场景有人脸识别以及其他需要人类输入信息的任务，这类任务中参与合作的人类扮演着提供输入的角色，来协助智能体完成任务；3）与前两种存在主导的合作方式不同，双向合作则指人类与智能体完成任务的动机与目标是相同的，人与智能体会相互向对方传递信息并进行行动配合以完成任务，人与智能体均会根据双方的状态调整自己的行为且人类与智能体都会影响环境，进而影响观测，常见的场景有游戏内的智

能体队友，辅助驾驶等。这类任务的特点是人与智能体的行为更为平等，决策上虽然独立，但存在内在的影响。我们可以认为双向合作是前两种类型在一定程度上的混合状态，双方均可在任意时刻发起沟通来传达自己的主导意图。

已有研究中，为了实现智能体对人的理解，通常使用逆强化学习方法（Inverse Reinforcement Learning，IRL）根据人类的专家数据来推断奖励函数以指导智能体学习。此外，在自然语言预训练大模型中，使用了基于人类反馈的强化学习（Reinforcement Learning from Human Feedback， RLHF），来对齐人类偏好，避免生成一些带有偏见的回复。但由于依赖大量真实人类数据，在一些危险或是复杂场景下难以获取足够的人类数据，无法实现实时的交互式人智协同。可解释人工智能（Explainable AI，XAI）则致力于让人类能够理解智能体的意图，帮助人类的心智模型向智能体的概念模型对齐，从而达到更好的人智协同。这两个类型的方法分别从智能体向人类学习和帮助人类的心智模型向智能体对齐来促进人与智能体的合作，可以认为是单向的对齐。近期，双向价值对齐（Bidirectional Value Alignment）的概念被提出，认为人智协同中的成功沟通可以通过双向价值对齐来表示，机器人准确地推断人类价值，并结合机器人对人类进行自身行为的有效解释。

但以上的这些价值对齐的方式与真正的双向人智协作还有一定的差距。在开放式世界的双向合作中，人类与智能体都是动态变化的，但上述的方法都将双方或其中一方视作是静态不变的。此外，在过往的研究中，强化学习常常将人类建模为玻尔兹曼模型，即将人视作是完全理性或至少是公平公正的，但这对于开放式环境而言是不够全面的。人在合作过程中的价值（或称意图/偏好）常常是会动态变化的，而由于智能体的动态适应变化，人类对智能体的认知，也是动态变化的。这要求智能体在合作中需要能够适应合作的人类的动态变化以及不同的策略意图，才能更好地配合人类。与此同时，人类也需要获取智能体的决策原因以及策略，适时调整自己的行为。**在这样的假设下，我们应当认为双向对齐是“从人类向智能体”与“从智能体向人类”这两个单向对齐所达成的动态平衡。**因此，有必要结合人机交互与多智能体强化学习，对人与智能体互动方式以及价值对齐进行更深入探索，研究能够动态适应人类价值变化的智能体，完成双向合作场景下的人智协同。

6. 总结

开放式多模态环境中的多智能体智能体学习是实现具备人类智能水平智能体的关键之一。一方面，现有的研究在竞争性环境场景下通过基于种群的学习在竞争型电子游戏取得重大突破。另一方面，随着预训练大模型的兴起，开始逐渐出现一些开放式环境的多模态基准数据集/任务出现，为相关研究打下基础，并在视觉导航等视觉-语言机器人控制研究中展现出一定效果。此外，在更多智能体交互类型和人智协同场景下，及更好地利用多模态信息，特别是自然语言的落实与通信，增强策略泛化性能，打通人机交互壁垒从而推动人智协同，也是未来需要积极探索的方向。

参考文献

[1] Yang Y, Luo J, Wen Y, et al. Diverse auto-curriculum is critical for successful real-world multiagent learning systems[J]. In Proc. of the 20th International Conference on Autonomous
Agents and Multiagent Systems, 2021, 1:51–56.

[2] Yuan L, Gao X, Zheng Z, et al. In situ bidirectional human-robot value alignment[J]. Science Robotics, 2022, 7(68): eabm4183.

[3] Hadfield-Menell D, Russell S J, Abbeel P, et al. Cooperative inverse reinforcement learning[J]. Advances in neural information processing systems, 2016, 29.

[4] Edmonds M, Gao F, Liu H, et al. A tale of two explanations: Enhancing human trust by explaining robot behavior[J]. Science Robotics, 2019, 4(37): eaay4663.

[5] Liu X, Jia H, Wen Y, et al. Towards Unifying Behavioral and Response Diversity for Open-ended Learning in Zero-sum Games[J]. Advances in Neural Information Processing Systems, 2021, 34: 941-952.

[6] Lindner D, El-Assady M. Humans are not Boltzmann Distributions: Challenges and Opportunities for Modelling Human Feedback and Interaction in Reinforcement Learning[J]. arXiv preprint arXiv:2206.13316, 2022.

[7] Sukhbaatar S, Fergus R. Learning multiagent communication with backpropagation[J]. Advances in neural information processing systems, 2016, 29.

[8] Rabinowitz N, Perbet F, Song F, et al. Machine theory of mind[C]//International conference on machine learning. PMLR, 2018: 4218-4227.

[9] Zhang K, Yang Z, Liu H, et al. Fully decentralized multi-agent reinforcement learning with networked agents[C]//International Conference on Machine Learning. PMLR, 2018: 5872-5881.

[10] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[11] Leibo J Z, Dueñez-Guzman E A, Vezhnevets A, et al. Scalable evaluation of multi-agent reinforcement learning with melting pot[C]//International Conference on Machine Learning. PMLR, 2021: 6187-6199.

[12] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[13] Ahn M, Brohan A, Brown N, et al. Do as i can, not as i say: Grounding language in robotic affordances[J]. arXiv preprint arXiv:2204.01691, 2022.

[14] Fan L, Wang G, Jiang Y, et al. MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge[J]. arXiv preprint arXiv:2206.08853, 2022.

[15] Reed S, Zolna K, Parisotto E, et al. A generalist agent[J]. arXiv preprint arXiv:2205.06175, 2022.

[16] Li Y, Zhang S, Sun J, et al. Cooperative Open-ended Learning Framework for

Zero-shot Coordination[J]. arXiv preprint arXiv:2302.04831, 2023.

[17] Bisk Y, Holtzman A, Thomason J, et al. Experience grounds language[J]. arXiv preprint arXiv:2004.10151, 2020.