兵棋推演的智能决策技术与挑战

doi: 10.16383/j.aas.c210547

Intelligent Decision Making Technology and Challenge of Wargame

- YIN Qi-Yue[1, 2],
- ZHAO Mei-Jing[1],
- NI Wan-Cheng[1, 2],
- ZHANG Jun-Ge[1, 2],
- HUANG Kai-Qi[1, 2]
- 1. Institute of Automation, Chinese Academy of Sciences, Beijing 100190
- 2. University of Chinese Academy of Sciences, Beijing 100049

Funds: Supported by National Natural Science Foundation of China (61906197)

More Information

- 人机对抗，作为人工智能技术的试金石，近年来获得了举世瞩目的进展. 随着 Deep Blue[1]、AlphaGo[2]、Libratus[3]、AlphaStar[4]等智能体分别在国际象棋、围棋、二人无限注德州扑克以及星际争霸中战胜顶尖职业人类选手，其背后的智能决策技术获得了广泛的关注，也代表了智能决策技术在中等复杂度完美信息博弈、高复杂度完美信息博弈再到高复杂度不完美信息博弈中的技术突破.

国际象棋、围棋代表了完美信息博弈，其状态空间复杂度由 10^{47} 增至 10^{360}，后者更是被誉为人工智能技术的阿波罗. 相比于上述两种博弈环境，二人无限注德州扑克，尽管状态空间复杂度仅有 10^{160}，但其为不完美信息博弈，相比于国际象棋与围棋信息集大小仅为 1，其信息集平均大小达到 10^{3}. 而星际争霸，作为高复杂度不完美信息博弈的代表，因其相比于上述游戏的即时制、长时决策等特性[4, 5]，对智能决策技术提出了更高的要求.

星际争霸突破之后，研究人员迫切需要新的人机对抗环境实现智能技术的前沿探索. 兵棋推演是一款经典策略游戏[6-8]，也被称为战争游戏，作为一种人机对抗策略验证环境，由于其具有不对称环境决策、更接近真实环境的随机性与高风险决策等特点，受到智能决策技术研究者的广泛关注. 近些年来，研究者投入了大量的精

力进行兵棋推演智能体研发以及兵棋推演子问题求解，试图解决兵棋推演的人机对抗挑战[9-14].

兵棋推演，一直以来都是战争研究和训练的手段，分为早期的手工兵棋与 20 世纪 80 年代后期普及的计算机兵棋[15-17]. 胡晓峰等人[6]全面综述了兵棋推演的基本要素(参演人员、兵棋系统模拟的战场环境和作战部队、导演部及导调机构)，指出“兵棋推演的难点在于模拟人的智能行为”，进而得出“兵棋推演需要突破作战态势智能认知瓶颈”，最后给出了如何实现态势理解与自主决策可能的路径. 和目前兵棋推演关注的重点不同，本文关注的是兵棋推演中的智能体研究，针对通用性的智能决策技术与挑战展开. 另外，需要阐明的是，本文中的兵棋推演，如非特别阐述，在不引起歧义的前提下统一指双方计算机兵棋推演(红蓝两方).

本文内容组织如下：第二章将梳理兵棋推演与目前主流人机对抗环境如星际争霸等的区别，以及为什么其潜在是人机对抗的下一个挑战；第三章将介绍兵棋推演智能技术的研究现状；之后在第四章阐述当前主流技术的瓶颈；第五章对兵棋推演的智能决策技术进行展望与思考，希望启发新的研究方向；最后对全文进行总结.

1. 兵棋智能决策问题的挑战

本章首先简要介绍兵棋推演问题以及与手工兵棋的比较. 在此基础上，以人机对抗发展脉络为主线，以兵棋推演中的智能体研究为核心，介绍兵棋推演与其他主流策略游戏的通用挑战，之后重点阐述兵棋推演的独特挑战. 前者为实现兵棋推演人机对抗的成功提供了技术基础，后者则对当下人机对抗智能体决策技术提出了新的挑战.

1.1 兵棋推演问题

早期的兵棋推演一般指手工兵棋，具有 200 年的研究历史，而随着信息技术与计算机性能的不断发展，计算机兵棋，因其简便、快速、逼真等特点成为目前兵棋推演的主流方向[18]. 王桂起等人[15]在 2012 年概述了兵棋的概念、发展、分类以及应用，并分析了兵棋的各组成要素以及国内外兵棋的研究现状. 彭春光等人[16]在 2009 年对兵棋推演技术进行了综述，指出兵棋主要研究人员决策与兵棋事件之间的因果关系.

2017 年，胡晓峰等人[6]对兵棋推演进行了全面的综述，描述了兵棋推演的基本要素，重点阐述了兵棋推演的关键在于模拟人的智能行为，面临的难点为“假变真”、“粗变细”、“死变活”、“静变动”、“无变有”，归结起来为“对战场态势的判断理解”以及“对未来行动的正确决策处置”，在此基础上，作者展望了 AlphaGo 等技术对兵棋推演带来的新机遇. 不同于上述工作，本文以人机对抗智能决策切入，针对通用性的智能决策技术与挑战展开对兵棋推演中的智能体研究.

1.2 策略游戏普遍挑战问题

回顾当前典型的已获得一定人机对抗突破的决策环境如雅达利、围棋、德州扑克以及星际争霸，可以得出一些基本的结论. 人机对抗研究的重心已经从早期的单智能体决策环境如雅达利过渡到了多智能体决策环境如围棋与星际争霸；从回合制决策环境如围棋逐渐过渡到更贴近现实应用的复杂即时战略类决策环境如星际争霸；从完美信息博弈如围棋逐渐过渡到非完美信息博弈如德扑与星际争霸；从以树为基础的博弈算法如围棋与德扑过渡到以深度强化学习为基础的大规模机器学习算法. 针对上述转变与各自博弈对抗环境的特点，可以凝练抽取一些影响智能体设计与训练的关键因素，如表 1 所述. 典型的兵棋推演仿真环境一般由算子、地图、想定以及规则要素组成，展现了红蓝双方之间的博弈对抗. 与代表性策略游戏如雅达利、围棋、德州扑克以及星际争霸等类似，兵棋推演的智能体研究表现出策略游戏中智能体研究的普遍挑战性问题.

表 1 对决策带来挑战的代表性因素

Table 1 Representative factors that challenge decision-making

游戏	雅达利	围棋	德州扑克	星际争霸	兵棋推演
不完美信息	√	×	√	√	√
长时决策	√	√	×	√	√
策略非传递	×	√	√	√	√
智能体协作	×	×	×	√	√
非对称环境	×	×	×	×	√
高随机性	×	×	×	×	√

不完美信息博弈. 不完美信息博弈是指没有参与者能够获得其他参与者的行动信息[19]，即参与者做决策时不知道或者不完全知道自己所处的决策位置. 相比于完美信息博弈，不完信息博弈挑战更大，因为对于给定决策点，最优策略的制定不仅仅与当下所处的子博弈相关. 与德州扑克、星际争霸相似，兵棋推演同样是不完美信息博弈，红方或者蓝方受限于算子视野范围、通视规则、掩蔽规则等，需要推断对手的决策进而制定自己的策略.

长时决策. 相比于决策者仅做一次决策的单阶段决策游戏，上述游戏属于序贯决策游戏[20]. 以围棋为例，决策者平均决策次数在 150 次，相比于围棋，星际争霸与兵棋推演的决策次数以千为单位. 长时决策往往导致决策点数量指数级的增加，使得策略空间复杂度变大，过高的策略空间复杂度将带来探索与利用等一系列难题，这对决策制定带来了极大的挑战.

策略非传递性. 对于任何策略 v_t 可战胜 v_{t-1}, v_{t+1} 可战胜 v_t，有 v_{t+1} 可战胜 v_{t-1}，则认为策略之间存在传递性. 一般情况下，尽管部分决策环境存在必胜策略，但在整个策略空间下都或多或少存在非传递性的部分，即大多数博弈的策略不具备传递性[21]. 例如，星际争霸与兵棋推演环境，策略难以枚举且存在一定的相互克制关系. 策略非传递性会导致标准自博弈等技术手段难以实现智能体能力的迭代提升，

而当前经典的博弈算法如 Double Oracle[58]等又往往难以处理大规模的博弈问题，使得逼近纳什均衡策略极其困难.

智能体协作. 在多智能体合作环境中，智能体间的协作将提升单个智能体的能力，增加系统的鲁棒性，适用于现实复杂的应用场景[22-24]. 围棋与两人德州扑克参与方属于纯竞争博弈环境，因此不存在多个智能体之间的协作. 星际争霸与兵棋虽然也属于竞争博弈环境，但是需要多兵力/算子之间配合获得多样化且高水平策略. 将上述问题看作是单个智能体进行建模对求解是困难的，可以建模为组队零和博弈，队伍之间智能体相互协作，最大化集体收益. 针对组队零和博弈问题，相比于二人零和博弈问题，理论相对匮乏.

为应对上述挑战，研究人员进行了大量的技术创新. 例如，在蒙特卡洛树搜索基础上引入深度神经网络实现博弈树剪枝、通过自博弈实现强化学习的围棋 AI AlphaGo 系列[2]，在虚拟遗憾最小化算法基础上引入安全嵌套子博弈求解以及问题约简等技术的二人无限注德州扑克 AI Libratus[3]，采用改进自博弈以及分布式强化学习的星际争霸 AI AlphaStar[4]. 上述技术为相应决策问题的挑战性因素提出了可行的解决方案，尽管兵棋推演存在上述挑战，但相关技术基础已经具备，可以指导兵棋推演的研究方向.

1.3 兵棋推演独特挑战问题

1.3.1 非对称环境决策

传统的非对称信息指某些行为人拥有但另一些行为人不拥有的信息，本文的非对称以学习的角度考虑，指的是游戏双方的能力水平或游戏平衡性. 以围棋、星际争霸以及绝大多数游戏环境为例，游戏设计者为保证游戏的体验以及促进人类选手竞技水平的提升，往往保证游戏不同方具有相对均衡的能力. 例如，星际争霸游戏中包含了三个种族，即人族、虫族以及神族，尽管不同种族具有截然不同的科技树、兵力类型等，但是三个种族在能力上处于大致均衡的状态.

相比于星际争霸等，兵棋推演中游戏是不平衡的. 这不仅体现在红方与蓝方在兵力配备上的不同，也体现在不同任务/想定下红方和蓝方的现实需要. 以部分夺控战为例，红方兵力水平一般弱于蓝方，同时红方往往具有更好的视野能力(如红方配备巡飞弹算子)，而蓝方往往具有更强的进攻能力(如配备更多的坦克算子). 这种严重的非对称性，对于目前的学习算法提出了极大的挑战.

当前主流的或改进的自博弈技术，在智能体迭代过程中往往对每个参与智能体以对称的方式进行训练，进而保证智能体能力在相互对抗的迭代过程中持续增长. 但是，在兵棋推演中，红方与蓝方严重的非对称性，使得直接采用相似的设计难以保证弱势方的训练，需要设计更合理的迭代方式(如启发式迭代)保证相对较弱势方的训练. 另一方面，在二人零和博弈中，虽然弱势方的纳什均衡策略可取，但是如何

根据对手的情况调整自己的策略以最大可能剥削或者发现对手的漏洞并加以利用，可能是要考虑的重点问题.

1.3.2 随机性与高风险决策

随机性与高风险主要体现在游戏的裁决中，泛指交战规则中随机影响因素以及对交战结果产生的影响. 裁决是游戏的重要组成部分，在决定游戏的胜负规则之外，明确定义了参与方在对抗过程中的交战结果. 例如，在围棋中，黑子包围白子之后，需要将白子从棋盘中拿下，即吃子. 在星际争霸环境中，两队兵力对抗中，血量为零的兵力将直接消失. 一般来说，在围棋等棋类游戏中，裁决不受随机因素的干扰，即不具有随机性. 而在星际争霸环境中，尽管不同兵力攻击产生的伤害数值是固定的，但仍然受到少量随机因素的影响，例如具有一定概率触发某项技能(如闪避).

相比于上述游戏，兵棋推演在所有攻击裁决过程中均受到随机因素的影响，即随机性较高，这主要是因为兵棋裁决一般遵循着“攻击等级确定、攻击等级修正、原始战果查询、最终战果修正”的基本流程. 在原始战果查询与最终战果修正中，将基于骰子产生的随机数值(两个骰子 1−12 点)分别进行修正，上述修正的结果差距较大，可能产生压制甚至消灭对方班组的战果，也有可能不产生任何效果. 更重要的是，相比于其他即时战略类游戏(如星际争霸)，兵力一旦消失，将不能重新生成，因此会造成极高的风险，对于专业级选手，兵力的消失往往意味着游戏的失败.

兵棋推演的随机性与高风险决策对于智能体的训练提出了极高的挑战. 反映在数据上，环境的状态转移不仅受到其他算子以及不可见信息的影响，也受到裁决的影响，即状态转移高度不确定. 另一方面，决策的高风险使得算子所处状态的值估计等具有高方差特性，难以引导智能体的训练，尤其是在评估上难以消除该随机性的情况下训练更加困难.

总的来说，兵棋推演，作为一种人机对抗策略验证环境，由于其具备目前主流对抗环境的挑战性问题，使得完成兵棋推演的人机对抗挑战具备一定的技术基础. 同时，由于其不对称信息决策、更接近于真实环境的随机性与高风险决策特点，对当前人机对抗技术提出了新的挑战，也使得兵棋推演成为人机对抗的下一个挑战.

2. 兵棋智能决策技术研究现状

为应对兵棋推演的挑战性问题，研究者提出了多种智能体研发与评测方法. 与围棋、星际争霸等主流游戏人机对抗智能体研发脉络类似(如星际争霸从早期知识规则为主，中期以数据学习为主，后期以联合知识与强化学习完成突破)，兵棋推演也经历了以知识驱动为主、以数据驱动为主以及以知识与数据混合驱动的研发历程. 兵棋的评测技术包含了智能体的定量与定性分析方法. 在本节中，将重点阐述兵棋智能体研发的技术与框架，同时对智能体的评估评测进行简述.

2.1 兵棋智能体研发技术与框架

当前智能体的研发技术与框架主要包含三类，即知识驱动、数据驱动以及知识与数据混合驱动的兵棋推演智能体，本节将分别阐述各个技术框架的研究进展.

2.1.1 知识驱动的兵棋推演智能体

知识驱动的兵棋推演智能体研发利用人类推演经验形成知识库，进而实现给定状态下的智能体决策[25]. 代表性的知识驱动框架为包以德循环(OODA[26])，其基本观点是通过观察(Observation)、判断(Orientation)、决策(Decision)以及执行(Action)的循环过程实现决策，如图 1 所示. 具体来说，观察包括观察自己、环境以及对手实现信息的收集；判断对应态势感知，即对收集的数据进行分析、归纳以及总结获得当前的态与势；决策对应策略的制定，利用前面两步的结果实现最优策略的制定；执行对应于具体的行动.

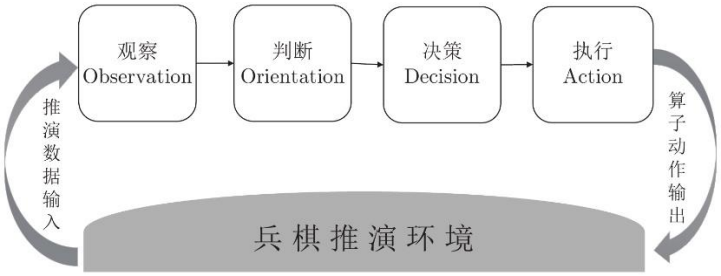

图 1 包以德循环

Fig. 1 OODA loop

通过引入高水平人类选手的经验形成知识库，可以一定程度规避前面所述的挑战性问题，实现态势到决策的规则制定与编码. 自 2017 年国内各类兵棋大赛举办以来，每年都有数十甚至上百个参赛队伍进行机机对抗，角逐的精英智能体将参与人机对抗以及人机混合对抗. 为适应不同的想定以及进行人机协同，目前绝大多数智能体为知识驱动型，即依据人类选手的经验进行战法总结，以行为树[27]、自动机[28]等框架实现智能体决策执行逻辑的编程实现. 总的来说，知识驱动型智能体研发依赖于人类推演经验与规律的总结，实现相对简单，不需要借助于大量的数据进行策略的训练与学习.

近些年来，通过编码高水平选手的决策，涌现出了一系列高水平知识驱动型智能体并开放对抗①，例如，信息工程大学的“兵棋分队级 AI-微风 1.0”，该智能体基于动态行为树框架，在不同想定下实现了不同的战法战术库. 中国科学院自动化研究所的“兵棋群队级 AI-紫冬智剑 2.0”，该智能体以 OODA 环为基本体系架构，以敌情、我情以及地形等通用态势认识抽象状态空间，以多层级任务行为认知抽象决策空间，可以快速适应不同的任务/想定. 目前部分智能体可以支撑人机混合对抗，甚至在特定想定下达到了专业级选手水平.

2.1.2 数据驱动的兵棋推演智能体

随着 AlphaGo、AlphaStar 等智能体取得巨大成功，以深度强化学习为基础进行策略自主迭代(如自博弈中每一轮的策略学习)成为当前的主流决策技术[29]并被成功应用于兵棋推演[30, 31]. 其基本框架如图 2 所示，智能体以自博弈或改进的自博弈方式进行每一代智能体的迭代，而每一代智能体采用强化学习的方式进行训练. 对于强化学习来说，智能体与环境进行交互收集状态、动作与奖赏等序列数据进行训练，直至学习得到可以适应特定任务的策略. 由于兵棋推演环境没有显式定义状态、动作与奖赏等的具体表现形式，因此在应用于强化学习的过程中，首要的任务是进行上述基本要素的封装，在此基础上便可以进行基本的强化学习训练.

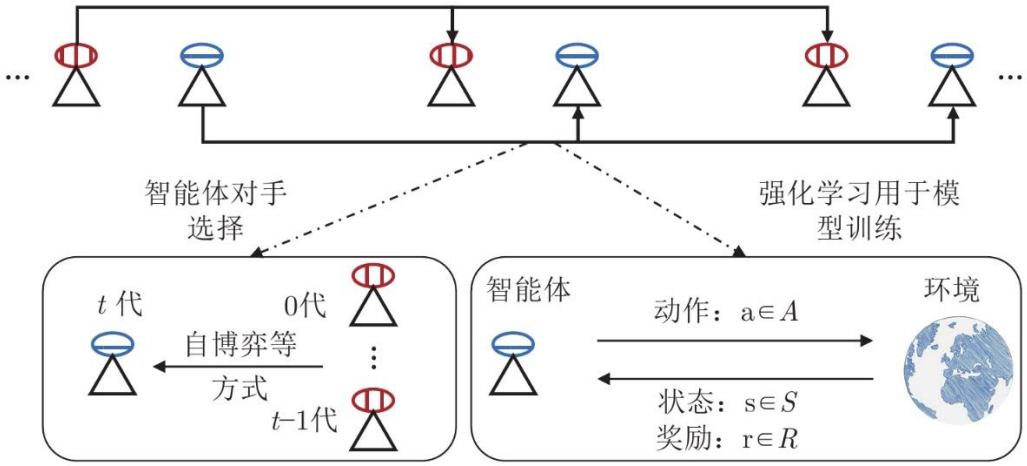

图 2 自博弈+强化学习训练

Fig. 2 Self-training + reinforcement learning

深度强化学习通过改进神经网络的设计可以一定程度缓解非完美信息与长时决策带来的挑战. 例如，通过增加认知网络结构如记忆单元[32, 33]可以有效使用历史信息，一定程度解决部分可观测状态下的决策问题；通过增加内在奖励驱动的环境建模网络[34]，可以缓解长时决策尤其是奖励稀疏情况下强化学习的训练. 自博弈尤其是改进的自博弈框架，如星际争霸提出的带有优先级的虚拟自我对局与联盟博弈有效缓解了策略非传递性的挑战，并通过初期的强化学习网络监督训练初始化实现了对策略非传递性的进一步缓解. 针对智能体协作，研究者提出了大量的多智能体协同算法，并通过奖励共享、奖励分配等实现了不同智能体的有效训练. 关于非对称性与高随机性，据本文作者所了解，尚未有相关文献解决兵棋推演的上述挑战.

近些年来部分研究者将其他数据学习方式与强化学习进行结合以缓解端到端强化学习的困难. 例如，李琛等[30]将 Actor-Critic 框架引入兵棋推演并与规则结合进行智能体开发，在简化想定(对称的坦克加步战车对抗)上进行了验证. 张振等[31]将近端策略优化技术应用于智能体开发，并与监督学习结合在智能体预训练基础上进行优化，在简化想定(对称的两个坦克对抗)验证了策略的快速收敛. 中国科学院自动化研究所提出的 AlphaWar[2]引入监督学习与自博弈技术手段实现联合策略的学习，保证了智能体策略的多样性，一定程度缓解了兵棋推演的策略非传递性问题. 2020

年, AlphaWar 在与专业级选手对抗过程中通过了图灵测试, 展现了强化学习驱动型兵棋推演智能体的技术优势.

另一方面, 分布式强化学习作为一种能够有效利用大规模计算资源加速强化学习训练的手段, 目前已成为数据驱动智能体研发的关键技术, 研究者提出了一系列算法在保证数据高效利用的同时也保证了策略训练的稳定性. 例如, Mnih 等人[35]在 2016 年提出异步优势动作评价算法(Asynchronous advantage actor-critic), 实现了策略梯度算法的有效分布式训练. Horgan 等人[36]在 2018 年提出 APE-X 分布式强化学习算法, 对生成数据进行有效加权, 提升分布式深度 Q 网络(Deep Q Network)训练效果. Mnih 等人[37]在 2018 年提出 IMPALA 算法实现了离策略分布式强化学习, 在高效数据产生的同时也可以通过 V-Trace 算法进行离策略(off-policy)修正, 该技术被成功用于夺旗对抗[38]. Espeholt 等人[39]在 2019 年引入中心化模型统一前向, 进一步提升了 IMAPLA 的分布式训练能力, 并被应用于星际争霸 AlphaStar 的训练中. 考虑到 IMPALA 的高效以及方便部署, 以 IMPALA 为代表的分布式强化学习已经成为兵棋智能体训练的常用算法. IMPALA 的结构如下图 3 所示, 其实现可以方便地通过 TensorFlow③、Pytorch④或伯克利近期提出的 Ray[40] 框架完成.

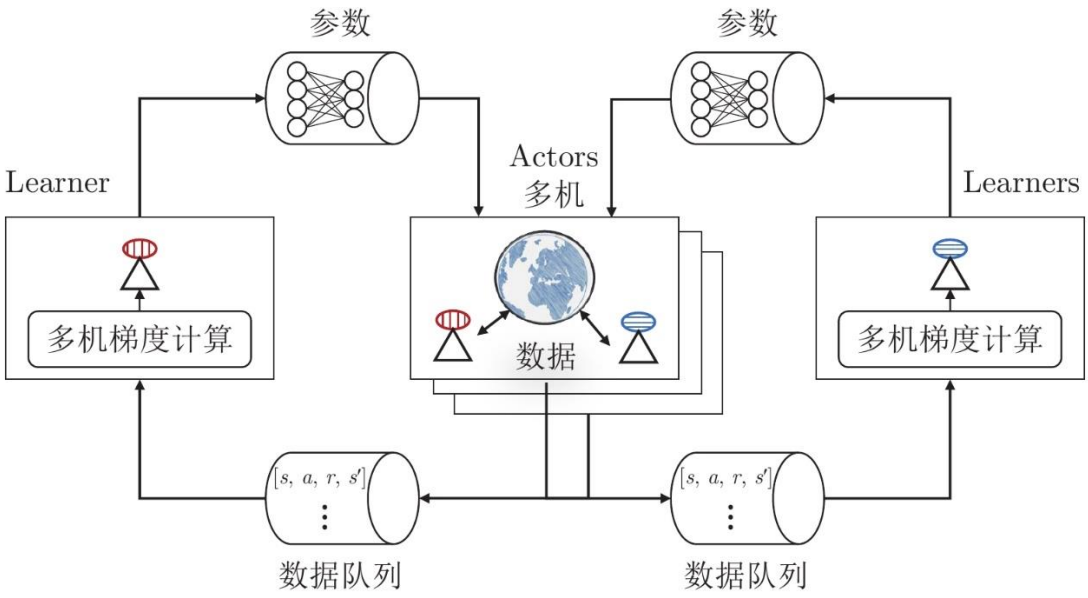

图 3 IMAPLA 用于兵棋推演 AI 训练

Fig. 3 IMAPLA for training wargame AI

2.1.3 知识与数据混合驱动的兵棋推演智能体

知识驱动智能体具有较强的可解释性, 但是受限于人类的推演水平. 与之相反, 基于数据驱动的兵棋智能体较少依赖人类推演经验, 可以通过自主学习的方式得到不同态势下的决策策略, 具有超越专业人类水平的潜力, 但是由于数据驱动的兵棋推演智能体依赖数据以及深度神经网络, 其训练往往较为困难且决策算法缺乏可解释性.

为了有效融合知识驱动与数据驱动框架的优点，避免各自的局限性，目前越来越多的研究者试图将两者进行结合[41]. 其中关注较多的工作为将先验信息加入到学习过程中进而实现对机器学习模型的增强[42-44]. 在该类工作中，知识或称为先验信息作为约束、损失函数等加入到学习的目标函数中实现一定程度的可解释性以及模型的增强. 近年来, Laura von Rueden 等人[42]进行了将知识融合到学习系统的综述并提出了知信机器学习的概念(informed machine learning)，从知识的来源、表示以及知识与机器学习管道的集成对现有方法进行了分类.

知识与数据混合驱动框架结合了两者的优势，可以更好应对兵棋推演环境的挑战，目前代表性的融合方式包括“加性融合”，如图 4 所示，即知识驱动与数据驱动各自做擅长的部分，将其整合形成完整的智能体. 一般来说，知识驱动善于处理兵棋推演前期排兵布阵问题，因为该阶段往往缺乏环境的有效奖励设计. 另一方面，紧急态势下的决策以及相对常识性的决策也可以由知识驱动完成，以减少模型训练的探索空间. 数据驱动善于自动分析态势并作出决策，更适用于进行兵棋推演中后期多样性策略的探索与学习. 此外，一些难以用相对有限的知识规则刻画的态势-决策也可由数据驱动完成. 黄凯奇等人[45] 提出了一种融合知识与数据的人机对抗框架，如图 5 所示，该框架以 OODA 为基础，刻画了决策不同阶段的关键问题，不同问题可以通过数据驱动或知识驱动的方式进行求解.

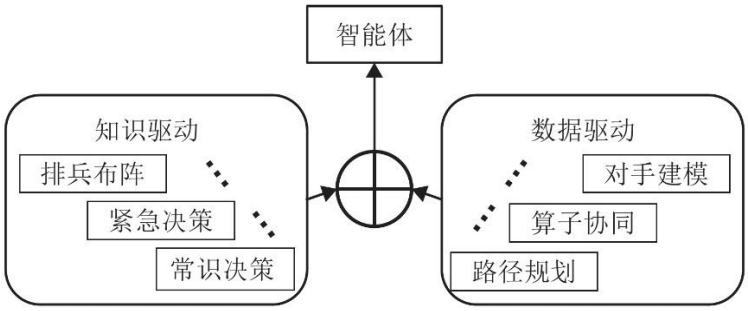

图 4 知识与数据驱动“加性融合”框架

Fig. 4 Additive fusion between knowledge-based and data-based AI

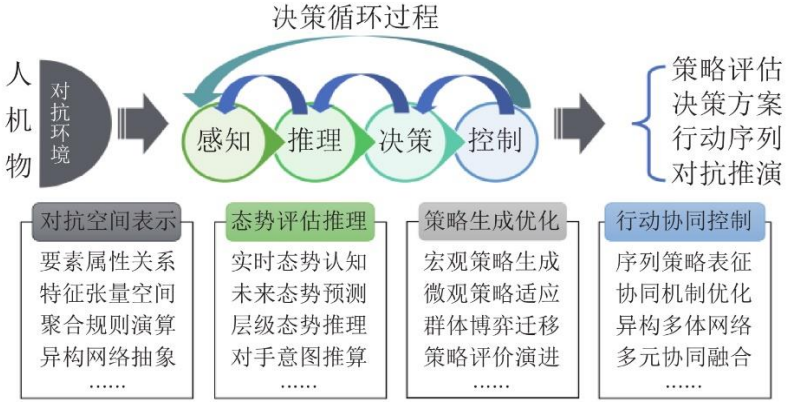

图 5 人机对抗框架[45]

Fig. 5 Human-machine confrontation framework[45]

另一种代表性融合方式为“主从融合”，如图 6 所示，即以一方为主要框架，另一方为辅助的融合方式. 在以知识驱动为主的框架中，整体设计遵循知识驱动的方式，在部分子问题或者子模块上采用如监督学习、进化学习等方式实现优化. 例如，武警警官学院开发的分队/群队 AI“破晓星辰 2.0”[5]在较为完善的人类策略库基础上结合蚁群或狼群等算法进行策略库优化，以提升智能体的适应性. 在以数据驱动为主的框架下，则采用如数据驱动的改进自博弈加强化学习的方式进行整体策略学习，同时增加先验尤其是常识性约束. 例如，将常识或人类经验作为神经网络选择动作的二次过滤以减少整体探索空间.

图 6 知识与数据驱动“主从融合”框架

Fig. 6 Hypotactic fusion between knowledge-based and data-based AI

2.2 兵棋智能体评估评测及平台

智能体的评估涉及智能体整体能力与局部能力评估，同时开放的智能体评估平台将有效支撑智能体的能力测评与迭代. 本节将从智能体评估算法与智能体评估开放平台展开介绍.

2.2.1 智能体评估算法

正确评估智能体策略的好坏对于智能体的训练与能力迭代具有至关重要的作用. 考虑到兵棋推演中策略的非传递性以及其巨大的策略空间问题，进行智能体的准确评估挑战巨大. 近年来，研究者们提出了一系列评估算法，试图对智能体能力进行准确描述. 经典的 ELO 算法[46]利用智能体之间的对抗结果，通过极大似然估计得到反映智能体能力的分值. 例如，围棋、星际争霸等对抗环境中的段位就是基于 ELO 算法计算获得. Herbrich 等人[47]提出 TrueSkill 算法，通过将对抗过程建立为因子关系图，借助于贝叶斯理论实现了多个智能体对抗中单一智能体能力的评估. 考虑到 ELO 算法难以处理策略非传递性这一问题, Balduzzi 等人[48]提出多维 ELO 算法，通过对非传递维度进行显式的近似改善了胜率的预测问题. 更进一步，Omidshafiei 等人[49]提出 α-rank 算法，基于 Markov-Conley 链，使用种群策略进化的方法，对多种群中的策略进行排序，实现策略的有效评估.

定量评估之外，也可以通过专家评判的方式进行定性评估，实现对智能体单项能力的有效评估. 例如，图 7 是庙算杯测试赛[6]中对智能体 AlphaWar 的评估，在人为抽象出的“武器使用”、“地形利用”、“兵力协同”、“策略高明”、“反应迅速”方面与测试赛排名第一位的人类选手进行了比较.

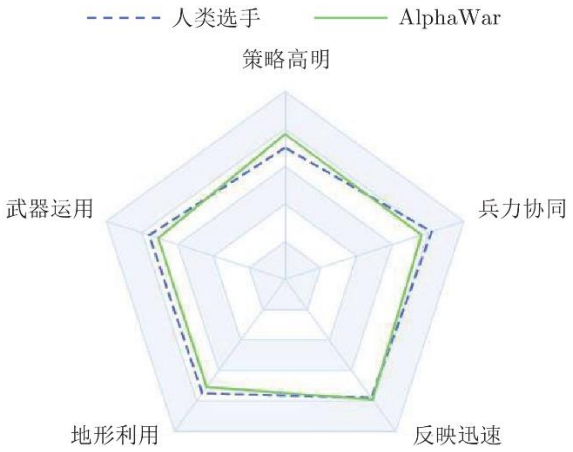

图 7 智能体单项能力评估

Fig. 7 Evaluation of specific capability of Agents

2.2.2 智能体评估开放平台

为促进兵棋推演智技术的发展，构建标准的评估评测平台至关重要，其可以实现广泛的兵棋智能体机机对抗、人机对抗甚至人机混合对抗[50]，这对兵棋推演评估评测平台提出了较高的要求，但也极大地促进了兵棋评估评测平台的建设与标准化. 最近，中国科学院自动化研究所构建了人机对抗智能门户网站 (http://turingai.ia.ac.cn/)，如图 8 所示. 该平台以机器和人类对抗为途径，以博弈学习等为核心技术来实现机器智能快速学习进化的研究方向. 平台提供兵棋推演智能体的机机对抗、人机对抗以及人机混合对抗测试，并支持智能体的多种评估评测.

图 8 “图灵网”平台

Fig. 8 Turing AI platform

3. 兵棋智能决策技术的挑战

针对兵棋推演的智能技术研究现状，本节重点阐述不同技术框架存在的挑战性问题，引导研究者对相关问题的深入研究.

3.1 知识驱动型技术挑战

知识驱动型作为智能体研发的主流技术之一，其依赖人类推演经验形成知识库，进而实现给定态势下的智能体决策. 基于此，知识驱动型智能体具有较强的可解释性，但同样面临不可避免的局限，即受限于人类本身的推演水平，同时环境迁移与适应能力较差，造成上述局限的根本原因在于缺乏高质量的知识库[51, 52]实现知识建模、表示与学习[53]，这也是目前知识驱动型技术的主要挑战. 知识库一般泛指专家系统设计所应用的规则集合，其中规则所联系的事实及数据的全体构成了知识库，其具有层次化基本结构.

对于兵棋推演来说，知识库最底层是“事实知识”，如算子机动能力等；中间层是用来控制“事实”的知识(规则、过程等表示)，对应于兵棋中的微操等；最顶层是“策略”，用于控制中间层知识，一般可以认为是规则的规则，如图 9 所示. 兵棋推演中知识库构建过程最大的挑战便是顶层策略的建模，面临着通用态势认知与推理困难的挑战. 胡晓峰等人[6]指出兵棋推演需要突破作战态势智能认知瓶颈，并提出战场态势层次不同，对态势认知的要求和内容也不同. 尽管部分学者尝试从多尺度表达模型[54]、指挥决策智能体认知行为建模框架[55]以及基于 OODA 环框架下态势认知概念模型[56]等进行态势建模，但是，目前基于经典知识规划的智能体受限于对环境的认识的正确性和完备程度，表现相较呆板缺乏灵活应对能力，不能很好地进行不确定环境边界下的意图估计与威胁评估等态势理解.

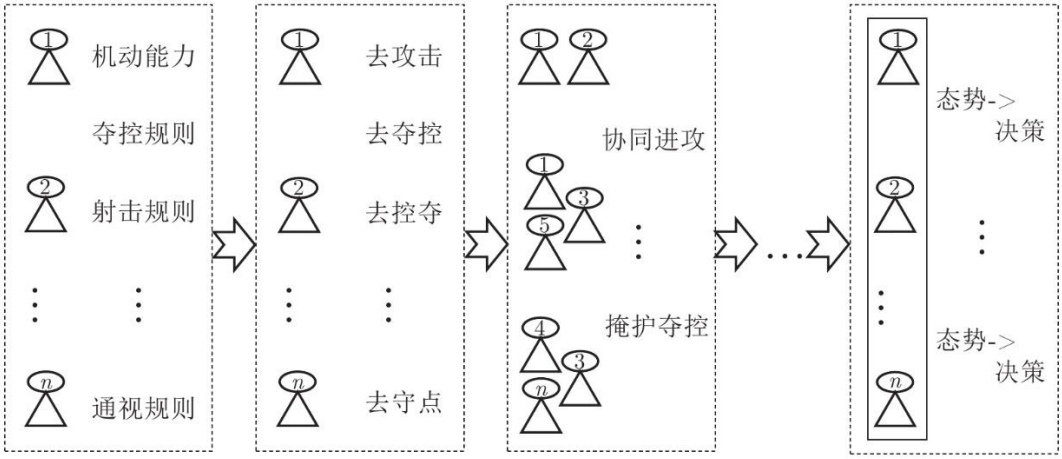

图 9 兵棋推演知识库构建示例

Fig. 9 Example of knowledge base construction for wargame

3.2 数据驱动型技术挑战

数据驱动型技术以深度强化学习为基础进行策略自主迭代，从该角度出发解决兵棋推演智能体研发，训练得到的智能体具有潜在的环境动态变化适应能力，甚至有可能超越专业人类选手的水平，涌现出新战法．同样地，为实现有效的智能体策略学习，目前数据驱动型技术面临以下技术挑战：自博弈与改进自博弈设计、多智能体有效协作、强化学习样本效率较低．其中，自博弈与改进自博弈设计可以实现智能体能力的有效迭代提升，多智能体有效协作将解决兵棋推演中的算子间协同(异步协同)问题，而解决强化学习样本效率较低问题可以实现在可控计算资源与时间下的智能体训练．

自博弈与改进自博弈．在兵棋推演这一二人零和博弈问题下，传统的博弈算法如虚拟自我对局[57]、Double Oracle[58]等难以适用于兵棋推演本身巨大的问题复杂度，采用目前较为主流的自博弈或改进自博弈方式实现智能体能力的迭代成为一种可行的方案．例如，围棋游戏的 AlphaGo 系列[2]采用结合蒙特卡洛树搜索与深度神经网络的自博弈强化学习实现智能体能力的迭代．星际争霸游戏的 AlphaStar[4]则改进传统的虚拟自我对局，提出带有优先级的虚拟自我对局并结合联盟博弈进行智能体迭代．具体来说，AlphaStar 引入主智能体、主利用智能体以及联盟利用智能体，并对不同的智能体采用不同的自博弈进行以强化学习为基础的参数更新．总的来说，尽管上述自博弈与一系列改进自博弈方法可以实现智能体的迭代，但当前的设计多是启发式迭代方式，兵棋推演的非对称环境等独特挑战是否适用有待验证与开展深入研究．

多智能体协作．协作环境下单个智能体的训练受到环境非平稳性的影响而变得不稳定[59-62]，研究者提出了大量的学习范式以缓解该问题，但仍然面临着智能体信用分配这一核心挑战，即团队智能体在和环境交互时产生的奖励如何按照各个智能体的贡献进行合理分配以促进协作[63-65]．目前，一类典型的算法为 Q 值分解类算法，即在联合 Q 值学习过程中按照单调性等基本假设将联合 Q 值分解为智能体 Q 值的联合，进而实现信用隐式分配[66-68]．例如，Sunehag 等人[66] 率先提出此类算法将联合 Q 值分解为各个智能体 Q 值的加和．在此基础上，Rashid 等人[67]基于单调性假设提出了更为复杂的 Q 值联合算法 QMIX．另外一类典型的信用分配算法借助于差异奖励(difference reward)来实现显式奖励分配．例如，Foster 等人[69]通过引入反事实的方法提出 COMA 以评估智能体的动作对联合智能体动作的贡献程度．通过将夏普利值引入 Q 学习过程中，Nguyen 等人[70]提出了 Shapley-Q 方法以实现“公平”的信用分配．在兵棋推演环境中，不同智能体原子动作执行耗时是不一样的，导致智能体协作时的动作异步性，如图 10 所示．这种异步性使得智能体间的信用分配算法要求的动作同步性假设难以满足，如何实现动作异步性下多智能体的有效协作仍然是相对开放的问题．

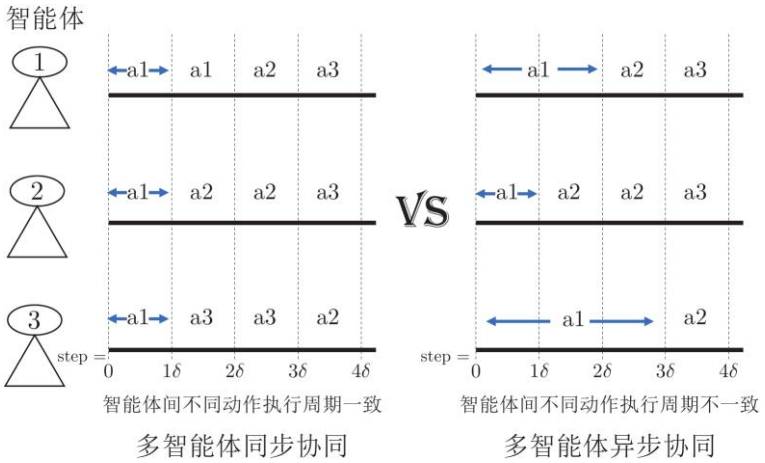

图 10 兵棋推演中的异步多智能体协同

Fig. 10 Asynchronous multi-agent cooperation in wargame

强化学习低样本效率. 强化学习通过与环境交互试错的方式进行模型训练, 一般样本效率较低, 因此在复杂环境下智能体训练需要动用巨大的计算资源. 例如, AlphaZero[71] 采用了 5000 一代 TPU 与 16 二代 TPU 进行智能体学习; AlphaStar[4]采用 192 TPU (8 核)、12 TPU (128 核)与 50400 CPU 实现群体博弈. 探索作为一种有效缓解样本效率低的手段[72], 近些年来受到了研究者的广泛关注, 并潜在适用具有巨大状态空间、稀疏奖励的兵棋推演环境中. 在单智能体强化学习中, 目前涌现了大量的探索类算法[72, 74], 如随机网络蒸馏(random network distillation)[34]、Go Explore[73] 等. 但多智能体的环境探索问题研究相对较少, 代表性方法包括 MAVEN[75]、Deep-Q-DPP[76]、ROMA[77] 等. 其中 MAVEN 通过在 QMIX 的基础上引入隐变量来实现多个联合 Q 值的学习, 进而完成环境的有效探索. Deep-Q-DPP 将量子物理中建模反费米子的行列式点过程(Determinantal Point Process)引入多智能体探索中, 通过增加智能体行为的多样性来实现探索. 另一方面, ROMA 通过考虑智能体的分工, 让相同角色的单元完成相似的任务, 进而利用动作空间划分来实现环境高效探索. 上述算法在星际争霸微操等验证环境中取得了有效的验证, 但是兵棋推演环境拥有更加庞大的状态空间, 如何实现智能体异步动作下的环境高效探索对当前技术提出了新的要求.

3.3 知识与数据混合驱动型技术挑战

知识与数据混合驱动型相比于知识型与数据型, 可以有效融合两者的优点, 既具备对环境的适用能力, 涌现出超越高水平人类玩家的策略, 同时又具备一定可解释性, 实现可信决策. 在融合过程中面临知识与数据驱动本身的技术挑战之外, 另一个核心技术挑战在于融合方式, 即如何实现两者的有机融合[78]. 上一章节提到了代表性的 “加性融合”、“主从融合”, 可以实现知识与数据的一定程度融合, 但是何种融合方式更优目前并无定论, 另一方面, 探索更优的兵棋推演知识与数据融合思路是值得深入探索与研究的开放问题.

加性融合的挑战. 在加性融合中, 知识驱动与数据驱动负责智能体不同的模块, 两者加和构成完整的智能体. 首先需要解决的问题是整个决策过程的模块化或解耦

合. 目前兵棋推演中较为简单的一种做法是开局过程(算子前期布局/机动到中心战场)采用知识驱动的方式, 中后期对抗(中心战场对抗如消灭对手、夺控等)采用数据驱动的方式. 但是上述做法如何解耦合或者定义两者的边界是困难的, 这不可避免引入专家的领域知识, 也将受限于专家对问题认识的局限. 以 OODA 为基础的人机对抗框架[45]虽然给出了较为一般化的框架, 但是如何在兵棋推演中具体实现存在较大的不确定性. 另一方面, 知识驱动与数据驱动部分相互制约, 在设计或训练过程中势必受到彼此的影响. 例如, 数据驱动的部分在迭代过程中受到知识驱动部分的限制. 这要求知识驱动或数据驱动部分在自我迭代的同时, 设计两者的交替迭代进而实现完整智能体能力的迭代提升. 上述设计与研究目前仍然是相对开放的问题.

主从融合的挑战. 在主从融合中, 以知识驱动或数据驱动为主, 部分子问题以另一种方式为手段进行解决. 在以数据驱动为主的框架中, 难点在于如何将知识或常识加入到深度学习或深度强化学习的训练中. 例如, 如何引入领域知识设计状态空间、动作空间以及奖赏. 相关设计将极大影响智能体的最终水平以及训练效率, 因此需要对上述问题进行折中, 保证智能体能力的同时尽可能引入更多的知识以提升训练效率. 在以知识为主的框架中, 难点在于寻找适宜用学习进行解决的子问题, 进而解决难以枚举或难以制定策略的场景. 例如采用经典的寻路算法[79]实现临机路障等环境下的智能体机动设计; 利用模糊系统方法实现兵棋进攻关键点推理[80]; 基于关联分析模型进行兵棋推演武器效用挖掘[81]. 目前, 在星际争霸、dota2 等复杂即时战略类游戏中的取得代表性成果的智能体多采用以数据驱动为主的方式, 即引入领域知识设计深度强化学习的各要素和训练过程, 如何根据兵棋推演独特的挑战进行相关技术迁移与改进目前是相对开放的问题.

3.4 评估评测技术挑战

当前智能体的评估主要借助机机对抗的胜率进行智能体综合能力/段位的排名/估计. 除此之外, 兵棋推演一般建模为多智能体协作问题, 因此, 单个智能体的能力评估将量化不同智能体的能力, 在人机协作[82]中机的能力评估中占据重要的地位. 另一方面, 人机对抗中人对机的主观评价正逐渐成为一种智能体能力评估的重要补充. 下面将分别介绍相关的挑战性问题.

非传递性策略综合评估. 多维 ELO 算法[48]在传统 ELO 的基础上通过对非传递维度进行显式的近似, 可以缓解非传递性策略胜率的预测问题, 但是因为其依赖于 ELO 的计算方式, 也就存在 ELO 本身对于对抗顺序依赖以及如何有效选取基准智能体等问题. 对于兵棋推演这一面临严重策略非传递性的问题, 目前的评估技术基于 ELO 或者改进的 ELO, 仍然具有较大的局限性.

智能体协作中的单个智能体评估. 基于经典的 ELO 算法, Jaderberg 等人[38]提出启发式的算法进行协作智能体中单个智能体的评估, 但是该算法依赖于智能体能力的可加和假设, 因此难以应用于兵棋推演环境, 即算子之间的能力并非线性可加和. 另一方面, TrueSkill 算法通过引入贝叶斯理论, 实现了群体对抗中的某一选手的评估,

但是其对时间不敏感，且往往会因为对抗选手的冗余出现评估偏差. 因此如何设计有效的评估算法实现协作智能体中的单个智能体的评估是当前的主要挑战之一.

定性评估标准体系化. 当前一些评估评测平台人为抽象了包括“武器使用”、“地形利用”等概念实现人机对抗中人对智能体的打分评测. 上述概念主要启发于指挥决策中对指挥官能力的刻画，因此是面向现实应用下智能体能力评估的重要维度[83, 84]. 但是，如何将智能体的评估体系与作战指挥中的能力维度进行对齐仍然是开放的问题，需要指挥控制领域的研究人员与博弈决策领域的研究人员共同协作.

4. 兵棋智能决策技术展望

为缓解兵棋推演智能决策技术存在的挑战性问题，部分研究者另辟蹊径，引入了新的理论、抽象约简问题等以应对兵棋推演的人机对抗.

4.1 兵棋推演与博弈理论

博弈理论是研究多个利己个体之间的策略性交互而发展的数学理论，作为个体之间决策的一般理论框架，有望为兵棋人机对抗挑战突破提供理论支撑[85-88]. 一般来说，利用博弈理论解决兵棋推演挑战，需要为兵棋推演问题定义博弈解，并对该解进行计算. 兵棋推演作为典型的两人零和博弈，可以采用纳什均衡解. 但是，纳什均衡解作为一种相对保守的解，并非在所有场合都适用. 考虑到兵棋推演的严重非对称性，纳什均衡解对于较弱势方可能并不合适. 因此，如何改进纳什均衡解(例如以纳什均衡解为基础进行对对手剥削解的迁移)是需要研究关键问题.

在博弈解求解这一问题上，早期相对成熟的求解方法包括线性规划、虚拟自我对局[57]、策略空间回应 oracle (Policy space response oracle)[89]、Double oracle[58]、反事实遗憾最小化[90] 等. 但是，上述纳什均衡解(或近似纳什均衡解)优化方法一般只能处理远低于兵棋推演复杂度的博弈环境，而目前主流的用于星际争霸等问题的基于启发式设计的改进自博弈迭代往往缺乏对纳什均衡解逼近的理论保证. 因此，针对兵棋推演这一具有高复杂度的不完美信息博弈问题，如何将深度强化学习技术有效地纳入可逼近纳什均衡解的计算框架、或者提出更有效/易迭代的均衡逼近框架，来实现兵棋推演解的计算仍然是开放性问题.

总的来说，尽管博弈理论为兵棋推演的人机对抗挑战提供了理论指导，但是，如何借助于该理论实现兵棋推演人机对抗的突破仍然是相对开放性的问题，需要研究者们进行更深入的研究.

4.2 兵棋推演与大模型

近些年来，大模型(预训练模型)在自然语言处理领域获得了飞速发展[91, 92]. 例如, OpenAI 于 2020 年发布的 GPT-3 模型参数规模达到 1750 亿[93]，可以作为有效的

零样本或小样本学习器提升自然语言处理下游任务的性能，如文本分类、对话生成、文本生成等. 中国科学院自动化研究所在 2021 世界人工智能大会上发布了三模态(视觉、文本、语音)大模型，具备跨模态理解与生成能力[7]. 一般来说，预训练的大模型，作为通用人工智能的一种有效探索路径，需要海量的数据支撑训练，但具有重要的学术研究价值与广阔的应用前景.

兵棋推演提供多种任务/想定，理论上可以有大量不同的训练环境，深度强化学习与环境交互试错的学习机制使得大模型训练的数据问题得以缓解. 但是，如何针对兵棋推演训练大模型，使得其在不同的兵棋对抗任务中可以快速适应仍然面临各种挑战，如图 11 所示. 首先，兵棋推演没有如自然语言处理任务较为通用的训练目标或优化目标，尤其是不同规模的对抗任务差异较大，因此如何设计该大模型的优化目标是需要解决的首要问题，这涉及强化学习中动作空间、奖励空间等多项要素的深入考虑.

图 11 兵棋推演大模型训练挑战

Fig. 11 Challenge of training big model for wargame

另一方面，兵棋推演包含异质且异步协同的智能体，不同任务下需要协同的智能体在数量、类型上有所差距，这就要求大模型在训练过程中既能解耦合不同智能体之间的训练，同时可以建立有效的协同机制实现智能体之间的协同. 尽管，可以采用智能体共享奖励、神经网络独立训练的框架，但是该设计过于简单，难以有效实现智能体协同时的信用分配等挑战性问题. 总的来说，如何设计大模型下多智能体训练以适应具有较大差异的兵棋推演任务是需要重点研究的问题之一.

最后，在自博弈过程中进行大模型的训练，需要适应不同规模(兵棋推演天然存在连队级、群队级、旅队级等规模)以及同规模下不同任务难度的对抗，这对大模型的训练提出了新的挑战. 自步学习[94]的范式提供了智能体由易到难的逐步训练框架，但如何定义兵棋推演不同任务难度是启发式的. 另一方面，要求智能体在更难任务训练时不能遗忘对已训练任务的记忆，这也需要持续学习[95]等前沿技术手段的引入.

4.3 兵棋推演关键问题抽象

星际争霸完整游戏的人机对抗挑战突破之前，研究者们设计了包括敌方意图识别[96]、微操控制(多智能体协同)[97, 98, 99]等在内的关键子任务以促进智能决策技术的发展. 针对兵棋推演问题，为引领技术突破进而反馈解决兵棋人机对抗挑战，迫切需

要对兵棋推演中的关键问题进行抽象、约简，在保证约简的问题能够表征原始问题的重要特征前提下，在约简的问题中进行求解.

基于上述考虑，本文提出两个约简问题，即排兵布阵与算子异步协同对抗. 需要指出的是，问题约简过程中不可避免对兵棋推演环境等要素的规则进行简化，甚至脱离兵棋推演本身的任务或者目的导向属性，但是相关问题的约简与抽象一定程度反映了兵棋推演智能体决策的核心挑战，将极大促进研究者对相关问题的研究.

排兵布阵. 排兵布阵反映了决策者在未知对手如何决策的前提下采取何种规划或者兵力选择可以最大化自己的收益，代表性环境如炉石传说卡牌类游戏，即如何布置自己的卡牌以在后期积累优势获得最大化利益. 其挑战在于未知对手如何规划的条件下实现己方规划，该问题因为缺乏验证环境，目前研究较少.

兵棋推演的前期，红方或者蓝方基于未知的对手信息布局自己的兵力，该布局一定程度决定了后期的对抗成败. 该过程因为缺少环境的显式反馈，无法度量何种排兵布阵能够最大限度利用地形、能够最大化攻击等，也就难以评估何种兵力布置最优. 基于上述原因，本文设计如图 12 所示的排兵布阵简化问题. 具体来说，在一个简化的地图中，红方与蓝方各占有一部分区域进行兵力放置，同时红方与蓝方之间具有一定距离间隔，考虑红方与蓝方不能移动且兵力放置之后自动进行裁决.

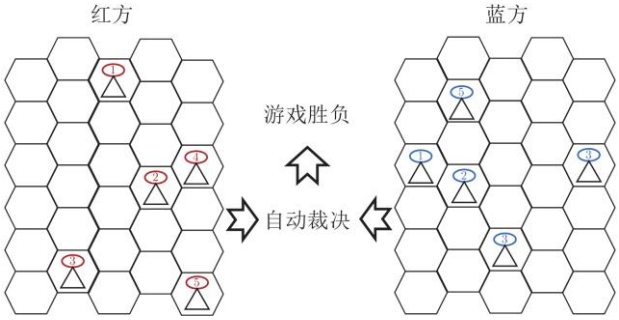

图 12 排兵布阵问题示意图

Fig. 12 Environment of arranging arms

需要指出的是，上述简化环境对兵棋推演本身做了极大的简化，更多是从算法研究的角度出发. 在研究兵力放置过程中，可以由简单到复杂进行调整，以契合兵棋推演问题本身，包括兵棋推演的目的加入(如夺控)、地形设置(如高程)等.

算子异步协同对抗. 算子协同对抗是多智能体相关问题的重要组成部分，目前相关领域已经开放了大量的智能体协同对抗环境，如星际争霸微操、捉迷藏等[22-24]. 值得注意的是，目前绝大多数环境，不同算子之间协同是同步的，即智能体的动作执行周期一致. 以此为基础，研究者提出了大量的算法实现有效的算子间协同[100, 75, 101]. 但是当不同智能体的动作执行周期不一致时，便导致异步协同问题，兵棋推演的对抗便属于异步协同对抗，当前的研究因为相关环境的缺乏相对较少.

兵棋推演中后期，红方与蓝方进行对抗，为评估智能体的接敌能力实现算子之间异步动作的有效协同，本文设计算子异步协同对抗简化问题. 如图 13 所示，在一

个简化的相对较小的地图上，不考虑复杂地形、复杂交战规则以及兵棋推演任务约束等因素，红方与蓝方在各自的起始位置出发进行对抗，算子可选动作包括机动(6个方向与停止)与射击(对方算子). 由于不同算子机动能力的差异，重点为领域提供多智能体异步协作的评估环境.

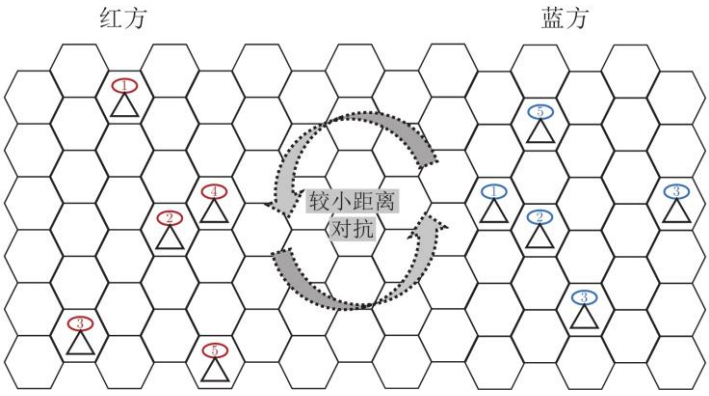

图 13 算子异步协同问题示意图

Fig. 13 Environment of asynchronous multi-agent cooperation

同排兵布阵问题，简化更多从验证算法性能的角度入手. 在研究算子异步协同对抗过程中，可以对任务的难度进行调整，如对地图进行调整，包括设置高程、增加特殊地形等.

为了促进上述问题的深入研究，在约简问题设计上，本文将陆续公开:

1)与 OpenAI Gym®一致的领域认可的环境接口，供智能体与环境交互进行策略的学习;

2)提供不同难度等级的内置智能体，供算法研究人员进行算法验证与算法间比较;

3)完全开放的底层源码，进而支持自博弈等主流技术以及人机对抗.

5. 结论

星际争霸人机对抗挑战的成功标志着智能决策技术在高复杂不完美信息博弈中的突破. 星际争霸之后，迫切需要新的人机对抗环境以牵引智能决策技术的发展. 兵棋推演，因其非对称信息决策以及随机性与高风险决策等挑战性问题，潜在成为下一个人机对抗热点. 本文详细分析了兵棋推演智能体的研究挑战尤其是其相比于其他博弈环境的独特挑战性问题，在此基础上了梳理了兵棋推演智能决策技术的研究现状，包括智能体研发技术框架以及智能体评估评测技术，之后指出了当前技术的挑战，并展望兵棋推演智能决策技术的发展趋势. 通过本文，将启发研究者对兵棋推演关键问题的研究，进而产生实际应用价值.

致谢

中国科学院自动化研究所的周雷博士在“兵棋推演与博弈理论”章节给出了博弈理论解决兵棋推演问题的研究思路，在此感谢周雷博士的建议.