智能博弈综述：游戏 AI 对作战推演的启示

孙宇祥[1]，彭益辉[1]，李斌[1]，周佳炜[1]，张鑫磊[1]，周献中[1,2]
（1. 南京大学工程管理学院，江苏 南京 210093；
2. 南京大学智能装备新技术研究中心，江苏 南京 210093）

摘　要：智能博弈领域已逐渐成为当前 AI 研究的热点之一，游戏 AI 领域、智能兵棋领域都在近年取得了一系列的研究突破。但是，游戏 AI 如何应用到实际的智能作战推演依然面临巨大的困难。综合分析智能博弈领域的国内外整体研究进展，详细剖析智能作战推演的主要属性需求，并结合当前最新的强化学习发展概况进行阐述。从智能博弈领域主流研究技术、相关智能决策技术、作战推演技术难点 3 个维度综合分析游戏 AI 发展为智能作战推演的可行性，最后给出未来智能作战推演的发展建议。以期为智能博弈领域的研究人员介绍一个比较清晰的发展现状并提供有价值的研究思路。

关键词：智能博弈；游戏 AI；智能作战推演；智能兵棋；深度强化学习
中图分类号：E91
文献标志码：A
doi: 10.11959/j.issn.2096−6652.202209

Overview of intelligent game: enlightenment of game AI to combat deduction

SUN Yuxiang[1], PENG Yihui[1], LI Bin[1], ZHOU Jiawei[1], ZHANG Xinlei[1], ZHOU Xianzhong[1,2]
1.School of Management and Engineering, Nanjing University, Nanjing 210093, China
2. Research Center for New Technology in Intelligent Equipment Nanjing University, Nanjing 210093, China

Abstract: The field of intelligent game has gradually become one of the hotspots of AI research. A series of research breakthroughs have been made in the field of game AI and intelligent wargame in recent years. However, how to develop game AI and apply it to the actual intelligent combat deduction is still facing great difficulties. The overall progress of research in the field of intelligent games in domestic and overseas were explored, the main attribute requirements of intelligent combat deduction was tracked, and it was summarized with the latest advancements in reinforcement learning. The feasibility of developing game AI into intelligent combat deduction were comprehensively analyzed from three dimensions: mainstream research technology in the field of intelligent game, relevant intelligent decision technology and technical difficulties of combat deduction, and finally, some suggestions for the development of future intelligent combat deductiongives were given. This paper can introduce a clear development status and provide valuable research ideas for researchers in the field of intelligent game.

Key words: intelligent game, game AI, intelligent combat deduction, intelligent wargame, deep reinforcement learning

0　引言

以 2016 年 AlphaGo 的成功研发为起点，对智能博弈领域的研究获得突飞猛进的进展。2016 年之前，对兵棋推演的研究还主要集中在基于事件驱动、规则驱动等比较固定的思路。到 2016 年，受 AlphaGo 的启发，研究人员发现智能兵棋、智能作战推演的实现并没有想象得那么遥远。随着机器学习技术的

收稿日期：2021–07–05；**修回日期：**2021–09–24
通信作者：周献中，zhouxz@nju.edu.cn
基金项目：国家自然科学基金资助项目（No.61876079）
Foundation Item: The National Natural Science Foundation of China (No.61876079)

发展，很多玩家十分憧憬游戏中有 AI 加入从而改善自己的游戏体验[1]。同时，在智能作战推演领域，不断发展的机器学习游戏 AI 技术也为智能作战推演的发展提供了可行思路[2]。传统作战推演 AI 主要以基于规则的 AI 和分层状态机的 AI 决策为主，同时以基于事件驱动的机制进行推演[3-4]。然而，随着近些年国内外在各种棋类、策略类游戏领域取得新突破，智能作战推演的发展迎来了新的机遇[5]。

国内游戏 AI 领域取得了标志性的进步。腾讯《王者荣耀》的《觉悟 AI》作为一款策略对抗游戏取得了显著成绩，可以击败 97%的玩家，并且多次击败顶尖职业团队[6]。网易伏羲人工智能实验室在很多游戏环境都进行了强化学习游戏 AI 的尝试[6]，如《潮人篮球》《逆水寒》《倩女幽魂》。超参数科技（深圳）有限公司打造了游戏 AI 平台“Delta”，集成机器学习、强化学习、大系统工程等技术，通过将 AI 与游戏场景结合，提供人工智能解决方案[7]。启元 AI“星际指挥官”在与职业选手的对抗中也取得了胜利[8]。北京字节跳动科技有限公司也收购了上海沐瞳科技有限公司和北京深极智能科技有限公司，准备在游戏 AI 领域发力。除了游戏 AI 领域，国内在智能兵棋推演领域也发展迅速。国防大学兵棋团队研制了战略、战役级兵棋系统，并分析了将人工智能特别是深度学习技术运用在兵棋系统上需要解决的问题[9]。中国科学院自动化研究所在 2017 年首次推出《CASIA-先知 1.0》兵棋推演人机对抗 AI[10]，并在近期上线“庙算 • 智胜”即时策略人机对抗平台[11]。此外，由中国指挥与控制学会和北京华戍防务技术有限公司共同推出的专业级兵棋《智戎 • 未来指挥官》在第三届、第四届全国兵棋推演大赛中成为官方指定平台。中国电科认知与智能技术重点实验室开发了 MaCA 智能博弈平台，也成功以此平台为基础举办了相关智能博弈赛事。南京大学、中国人民解放军陆军工程大学、中国电子科技集团公司第五十二研究所等相关单位也开发研制了具有自主知识产权的兵棋推演系统[12-15]。2020 年，国内举办了 4 次大型智能兵棋推演比赛，这些比赛对于国内智能博弈推演的发展、作战推演领域的推进具有积极影响。游戏 AI 和智能兵棋的发展也逐渐获得了国内学者的关注，胡晓峰等人[5]提出了从游戏博弈到作战指挥的决策差异，分析了将现有主流人工智能技术应用到战争对抗过程中的局限性。南京理工大学张振、李琛等人利用 PPO、A3C 算法实现了简易环境下的智能兵棋推演，取得了较好的智能性[16-17]。中国人民解放军陆军工程大学程恺、张可等人利用知识驱动及遗传模糊算法等提高了兵棋推演的智能性[18-19]。中国人民解放军海军研究院和中国科学院自动化研究所分别设计和开发了智能博弈对抗系统，对于国内智能兵棋推演系统的开发具有重要参考价值[20]。中国人民解放军国防科技大学刘忠教授团队利用深度强化学习技术在《墨子•未来指挥官系统》中进行了一系列智能博弈的研究，取得了突出的成果[21]。中国科学院大学人工智能学院倪晚成团队提出一种基于深度神经网络从复盘数据中学习战术机动策略模型的方法，对于智能博弈中的态势认知研究具有重要参考价值[22]。

总体来说，国内在智能博弈领域进行了一系列的研究，尝试将该技术应用到作战推演领域，建立了具有自主产权的博弈平台，技术层面也不断突破，不再局限于传统的行为决策树、专家知识库等，开始将强化学习技术、深度学习技术、遗传模糊算法等引入智能博弈，取得了一系列的关键技术的突破。但是，当前的研究主要聚焦在比较简单的智能博弈环境，对复杂环境及不完全信息的博弈对抗研究仍然需要进一步探索。

国外游戏 AI 领域则取得了一系列突出成果，尤其是深度强化学习技术的不断发展，游戏 AI 开始称霸各类型的游戏[23]。2015 年 DeepMind 团队发表了深度 Q 网络的文章，认为深度强化学习可以实现人类水平的控制[24]。2017 年，DeepMind 团队根据深度学习和策略搜索的方法推出了 AlphaGo[25]，击败了围棋世界冠军李世石。此后，基于深度强化学习的 AlphaGo Zero[26]在不需要人类经验的帮助下，经过短时间的训练就击败了 AlphaGo。2019 年，DeepMind 团队基于多智能体（agent）深度强化学习推出的 AlphaStar[27]在《星际争霸 II》游戏中达到了人类大师级的水平，并且在《星际争霸 II》的官方排名中超越了 99.8%的人类玩家。《Dota 2》AI“OpenAI Five”在电竞游戏中击败世界冠军[28]，Pluribus 在 6 人无限制德州扑克中击败人类职业选手[29]。同时 DeepMind 推出的 MuZero 在没有传授棋类运行规则的情况下，通过自我观察掌握围棋、国际象棋、将棋和雅达利（Atari）游戏[30]。与军事推演直接相关的《CMANO》和《战争游戏：红龙》（Wargame：Red Dragon），同样也结合了最新的机

器学习技术提升了其智能性[31]。美国兰德公司也对兵棋推演的应用进行相关研究，利用兵棋推演假设分析了俄罗斯和北大西洋公约组织之间的对抗结果，并利用智能兵棋推演去发现新的战术[32]。兰德研究员也提出将兵棋作为美国军事人员学习战术战法的工具[33]。美国海军研究院尝试使用深度强化学习技术开发能够在多种单元和地形类型的简单场景中学习最佳行为的人工智能代理，并将其应用到军事训练及军事演习[34-35]。

但就目前而言，国外的研究也遇到了瓶颈。虽然也尝试将深度强化学习技术利用到作战领域，但是就目前发表的论文和报告来看，国外学者、研究人员将机器学习技术应用到作战推演 AI 中还有很多问题需要解决，现阶段也是主要在游戏 AI 领域及简单的作战场景进行实验验证及分析。作战推演 AI 的设计也不仅仅是把机器学习技术照搬照用这么简单。但是必须肯定的是，随着未来计算机硬件的发展和机器学习技术的完善，作战推演 AI 会迎来一波革命式的发展，给各类作战智能指挥决策带来翻天覆地的变化。本文从智能博弈的重要应用领域——作战推演分析了国内外整体背景，进而引出作战推演的技术需求，并给出当前可参考的主流及小众技术思路。同时，对可能出现的技术难点进行了分析并给出解决方案建议。最后，对作战推演的未来发展提出建议。

1　智能作战推演主要属性需求

1.1　状态空间

状态空间是作战推演中的每个作战实体的位置坐标、所处环境、所处状态等要素的表现，是深度强化学习进行训练的基础。在围棋中，状态空间就是棋盘上每个点是否有棋子。在《觉悟 AI》中，状态空间是每一帧、每个单位可能有的状态，如生命值、级别、金币[36-39]。在《墨子 • 未来指挥官系统》中，状态空间主要是每个作战单元实体的状态信息，是由想定中敌我双方所有的作战单元信息汇聚形成的。本节尤其要明确状态空间和可观察空间是可区分的，可观察空间主要是每个 agent 可以观察到的状态信息，是整个状态空间的一部分。作战推演中的状态空间将更加复杂，具有更多的作战单位和单位状态。针对敌我双方的不同作战单位、不同单位属性、不同环境属性等定义作战推演的状态空间属性。例如敌我双方坦克单元应包括坐标、速度、朝向、载弹量、攻击武器、规模等。陆战环境应包括周围道路信息、城镇居民地、夺控点等。

1.2　动作空间设计

动作空间是指在策略对抗游戏中玩家控制算子或游戏单元可以进行的所有动作的集合。对于围棋来说，动作空间为 361 个，是棋盘上所有可以落子的点。对于《王者荣耀》和《Dota》这类游戏来说，动作空间主要是玩家控制一个“英雄”进行的一系列操作，玩家平均水平是每秒可以进行一个动作，但是需要结合走位、释放技能、查看资源信息等操作。例如《觉悟 AI》的玩家有几十个动作选项，包括 24 个方向的移动按钮和一些释放位置/方向的技能按钮[34]。因此每局多人在线战术竞技（multiplayer online battle arena，MOBA）游戏的动作空间可以达到 $10^{60\,000+}$。假设游戏时长为 45 min，每秒 30 帧，共计 81 000 帧，AI 每 4 帧进行一次操作，共计 20 250 次操作，这是游戏长度。任何时刻每个“英雄”可能的操作数是 170 000，但考虑到其中大部分是不可执行的（例如使用一个尚处于冷却状态的技能），平均的可执行动作数约为 1 000，即动作空间[37]。因此，操作序列空间约等于 $1\,000^{20\,250}=10^{60\,750}$。而对于《星际争霸》这类实时策略对抗游戏来说，因为需要控制大量的作战单元和建筑单元，动作空间可以达到 $10^{52\,000}$[38]。而对于《CMANO》和《墨子 • 未来指挥官系统》这类更加贴近军事作战推演的游戏来说，需要对每个作战单元进行大量精细的控制。在作战推演中，每个作战单元实际都包括大量的具体执行动作，以作战飞机为例，应包括飞行航向、飞行高度、飞行速度、自动开火距离、导弹齐射数量等。因此，实际作战推演需要考虑的动作空间可以达到 $10^{100\,000+}$。可以看出，对于作战推演来说，庞大的动作空间一直是游戏 AI 迈进实际作战推演的门槛。现有的解决思路主要是考虑利用宏观 AI 训练战略决策，根据战略决策构建一系列绑定的宏函数，进行动作脚本设计。这样的好处是有效降低了动作空间设计的复杂度，同时也方便高效训练，但是实际问题是训练出来的 AI 总体缺乏灵活性，过于僵化。

对于动作空间，还需要考虑其是离散的还是连续的，Atari 和围棋这类游戏动作都是离散动作空间[25,39-40]，《星际争霸》《CMANO》《墨子 • 未来指挥官系统》这类游戏主要是连续动作空间[38]。对于离散动作，可以考虑基于值函数的强化学习进行

训练，而对于连续动作，可以考虑利用基于策略函数的强化学习进行训练。同时，离散动作和连续动作也可以互相转化。国内某兵棋推演平台由原先的回合制改为时间连续推演，即把回合制转化为固定的时间表达。同时对于连续动作，也可以在固定节点提取对应的动作，然后将其转化为离散动作。

1.3 决策空间构建

智能博弈中的决策主要是指博弈对抗过程中的宏观战略的选择以及微观具体动作的选择。宏观战略的选择在《墨子·未来指挥官系统》推演平台中体现得比较明显。在推演比赛开始前，每个选手要进行任务规划，这个任务规划是开始推演前的整体战略部署，例如分配导弹打击目标，规划舰艇、战斗机活动的大致区域，以及各个任务的开始执行时间等。这一决策空间与想定中的作战单元数量、任务规划数量相关。在制定完成宏观战略决策后，推演阶段即自主执行所制定的宏观战略决策。同时，在推演过程中也可以进行微观具体动作的干预，这一阶段的具体动作和作战单元数量、作战单元动作空间成正比。在实际作战推演中利用智能算法进行智能决策，首先需要明确决策空间数量。在现有的《墨子·未来指挥官系统》中，针对大型对抗想定，计算机基本需要每秒进行数百个决策，一局想定推演中双方博弈决策空间数量预估为 10^{80+}个，而对于《星际争霸》《Dota 2》和《王者荣耀》这类即时战略（real-time strategy，RTS）游戏，决策空间会低一些。实际作战推演每小时的决策空间数量将高于 10^{50+}个。对于这类智能决策的方案，现有 RTS 游戏中提出的思路是利用分层强化学习的方法进行解决，根据具体对抗态势进行宏观战略决策的选择，然后根据不同的决策再分别执行对应的微观具体动作，这样可以有效降低智能决策数量，明显提高智能决策的执行效率。

1.4 胜利条件设置

博弈对抗的胜利是一局游戏结束的标志。而不同游戏中的胜利条件类型也不同，围棋、国际象棋这些棋类博弈对抗过程中有清晰明确的获胜条件[30]。而 Atari 这类游戏[40]只需要获得足够的分数即可获得胜利。对于《王者荣耀》这类推塔游戏，不管过程如何，只要最终攻破敌方水晶就可以获取胜利。这些胜利条件使得基于深度强化学习技术的游戏 AI 开发相对容易，在回报值设置中给予最终奖励更高的回报值，总归能训练出较好的 AI 智能。然而对于策略对抗游戏，甚至实际作战推演来说，获胜条件更加复杂，目标更多。比如，有时可能需要考虑在我方损失最低的情况下实现作战目标，而有时则需要不计代价地快速实现作战目标，这些复杂多元的获胜条件设置将使得强化学习的回报值设置不能是简单地根据专家经验进行赋值，而需要根据真实演习数据构建奖赏函数，通过逆强化学习技术满足复杂多变的作战场景中不同阶段、不同目标的作战要求。

1.5 回报值设置

博弈对抗过程中最核心的环节是设置回报值，合理有效的回报值可以保证高效地训练出高水平 AI。对于《星际争霸》《王者荣耀》等游戏，可以按照固定的条件设置明确的回报值，例如将取得最终胜利设置为固定的回报值。但是一局游戏的时间有时较长，在整局对抗过程中，如果只有最终的回报值将导致训练非常低效。这就是作战推演中遇到的一个难题，即回报值稀疏问题。为了解决这个难题，现有的解决方案都是在对抗过程中设置许多细节条件，如获得回报值或损失回报值的具体行为。比如在“庙算·智胜”平台中的博弈对抗，可以设置坦克击毁对手、占领夺控点即可获得回报值，如果被打击、失去夺控点等则会损失回报值，甚至为了加快收敛防止算子长期不能达到有效地点，会在每步（step）都损失微小的回报值。《觉悟 AI》也同样设置了详细的奖赏表[36]，从资源、KDA（杀人率（kill，K），死亡率（death，D），支援率（assista，A））、打击、推进、输赢 5 个维度设置了非常详细的具体动作回报值。这样就可以有效解决回报值稀疏的问题。但是，对于复杂的作战推演来说，设计回报函数可能还需要更多的细节。因为作战情况将更加复杂多样，需要利用逆强化学习[41-42]，通过以往的作战数据反向构建奖赏函数。

1.6 战争迷雾

战争迷雾主要是指在博弈对抗过程中存在信息的不完全情况，我方并不了解未探索的区域实际的态势信息。围棋、国际象棋这类博弈对抗游戏中不存在这类问题。但是在《星际争霸》《Dota 2》《王者荣耀》以及《CMANO》等 RTS 游戏中设计了这一机制。实际的作战推演过程中同样也存在此类问题，但是情况更加复杂。在实际作战推演中，可以考虑利用不完全信息博弈解决这个问题，已有学者利用不完全信息博弈解决了德州扑克中的不完全

信息问题[29]，但是在实际作战推演中这一问题还需要进一步探讨研究。

1.7　观察信息

这里需要对智能博弈中的观察信息与游戏状态空间进行区分，观察信息主要是指博弈的 agent 在当前态势下可以获取的态势信息，是部分状态信息。由于在智能博弈对抗过程中会产生战争迷雾问题，因此需要在处理博弈信息时设置 agent 可以获取到的信息。《星际争霸》中观察信息主要有两层意思，一个层面是屏幕限制的区域更易于获取态势信息，因为玩家更直观的注意力在屏幕局域，部分注意力在小地图局域。为了更加符合实际，AlphaStar 也按照这种限制对《星际争霸》中的注意力区域进行限制，从而更好地防止 AI 产生作弊行为。而这也是部分《星际争霸》AI 被人诟病的原因，即没有限制机器的关注区域。另一个层面是对《星际争霸》中作战单元可观察区域内的态势信息进行获取，对于不能获取的态势信息则只能评估预测，而这一部分则涉及对手建模部分，主要利用部分可观察马尔可夫决策过程（partially observable Markov decision process，POMDP）[43]，这一技术明显难于完全信息博弈。而对于围棋游戏来说，其中的态势信息是完全可获取的，属于完全信息博弈，态势信息即观察信息。并且围棋游戏属于回合制，相对于即时策略游戏，其有更加充分的获取态势信息的时间。因此，则可以利用蒙特卡洛树搜索（Monte Carlo tree search，MCTS）算法对所获取的围棋游戏中的观察信息进行详细分析，计算出所有可能的结果，进而得出最佳的方案策略。《Dota 2》中的观察信息是指所控制的某个“英雄”所获取的态势信息，其主要也是对主屏幕的态势信息和小地图的态势信息进行结合处理。《王者荣耀》也与此类似，其主要以小地图的宏观信息进行训练，然后以此为基础为战略方案提供支持，如游戏中的“英雄”是去野区发育还是去中路对抗。同时，对主屏幕态势信息进行特征提取，结合强化学习训练，可以得出战术层面的方案和建议，是去选择回塔防御还是进草丛躲避，或者推塔进攻。墨子兵棋推演系统和《CMANO》则更加接近真实作战推演，在作战信息获取各个方面都高度模拟了作战推演的场景，需要获取具体的对空雷达、对地雷达、导弹探测、舰艇雷达等信息后才能判断态势信息，这部分可观察信息非常复杂，需要结合各种情况才能发现部分目标，对于战争迷雾更加真实。因此，作战推演观察信息完全可以借鉴 POMDP 进行可观察信息建模，但还需要设置各种更加符合真实装备的作战情况，需要在环境中提前设置有针对性的条件。

1.8　对手建模

在博弈对抗过程中对手 AI 的建模也是至关重要的，不同水平的 AI 会导致博弈对抗的胜率不同，并且直接影响推演对抗的价值[39-45]。如果对手 AI 水平过低，就不能逼真地模拟假设对手，博弈过程和推演结果也价值不高。在 DeepMind 开发的 AlphaGo 和 AlphaStar 中，AI 性能已经可以击败职业选手，通过训练后产生的决策方案已经可以给职业选手新的战术启发。国内《墨子•未来指挥官系统》也与国内高校合作，研发的基于深度强化学习的智能 AI 已经可以击败全国兵棋大赛十强选手。而在中国科学院自动化研究所开发的“庙算•智胜”上，积分排名前三名的均是 AI 选手，胜率均在 80%以上[11]。但是，现有对手建模主要还是聚焦在一对一的对手建模，很少学者研究多方博弈，而这在实际作战推演中更加需要。在实际作战对抗博弈过程中普遍会考虑多方博弈，如在《墨子•未来指挥官系统》的海峡大潮想定中，红方不仅面对蓝方，还有绿方，蓝方和绿方属于联盟关系。这就需要在对手建模中充分考虑这种复杂的博弈关系。

1.9　想定设计

博弈对抗的环境因素也是影响智能决策的重要因素之一。在围棋、国际象棋这些棋类游戏中，想定是永久固定不变的，而且也完全没有环境的影响，因此 AlphaGo 这类智能 AI 完全没有考虑环境的因素。在《觉悟 AI》《Dota 2》这类游戏中就需要考虑不同“英雄”在同一个场景中会产生不同的影响。不同的“英雄”阵容搭配也会对推演结果产生不同的影响，《觉悟 AI》尝试利用强化学习技术，结合历史数据解决这一问题。这对于作战推演的武器装备搭配也具有启发价值。但是在实时策略游戏中要考虑更加复杂的环境因素及其影响，不仅作战单元会产生变化，并且在不同的作战推演中，不同的环境之中也会有不同的地形、地貌，这些因素会对作战推演的过程产生非常重要的影响。《CMANO》《墨子 • 未来指挥官系统》《战争游戏：红龙》中都需要考虑地形因素。例如《CMANO》中登陆作战需要考虑水雷所在区域、登陆舰艇吃水深度，否则会产生搁浅，不能在理想区域登陆会对

作战目标产生较大负面影响。因此，对于实际作战推演来说，最大的挑战是防止训练的深度强化学习 AI 对某个想定产生过拟合。作战场景是千变万化的，传统的基于规则的 AI 就很难适应变化的想定，早期的《先知•兵圣》比赛中就比较突出地显示了这一问题。强化学习也容易训练出某个过拟合的模型，导致只在某个想定会有较好的 AI 智能性，假如更换作战想定就需要重新训练很长时间。为了解决这一问题，现有思路是利用迁移学习、先验知识和强化学习的思路来增强算法的适应性，并可以加速回报函数收敛，保证快速训练出高水平的 AI 模型。

1.10　总体比较

本节针对智能作战推演所需要的关键属性，结合当前游戏 AI、智能兵棋等相关博弈平台，利用相关文献[6,8,24-25,29-30,37-39,43,46-49]进行分析，经过对比不难发现游戏 AI 过渡到智能兵棋，甚至是智能作战推演的难度，各个关键属性也是未来需要研究突破的关键点，具体见表 1。

2　作战推演的智能决策核心技术思路

2.1　强化学习技术框架

强化学习的核心思想是不断地在环境中探索试错，并通过得到的回报值来判定当前动作的好坏，从而训练出高水平的智能 AI[50]。马尔可夫决策过程（Markov decision process，MDP）是强化学习的基础模型，环境通过状态与动作建模，描述 agent 与环境的交互过程。一般地，MDP 可表示为四元组 $<S,A,R,T>$[44]：

- S 为有限状态空间（state space），包含 agent 在环境中的所有状态；
- A 为有限动作空间（action space），包含 agent 在每个状态上可以采取的所有动作；
- R 为奖赏函数（reward function），$R_{ss'}^{a}$ 表示 agent 在状态 s 下执行动作 a，到达下一状态 s'，从环境交互中获取的奖励；
- T 为环境的状态转移函数（state transition function），$P_{ss'}^{a}=P\left[S_{t+1}=s'\mid S_t=s,A_t=a\right]$ 表示在状态 s 下执行动作 a，并转移到状态 s' 的概率。

在 MDP 中，agent 与环境交互如图 1 所示

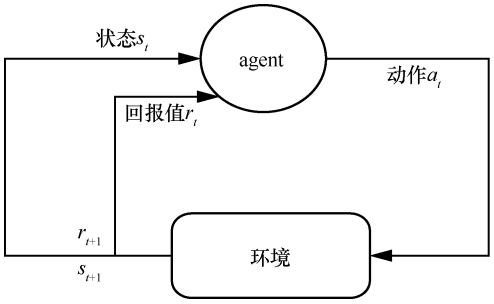

图 1　agent 与环境交互

agent 从环境中感知当前状态 s_t，从动作空间 A 中选择能够获取的动作 a_t；执行 a_t 后，环境给 agent 相应的奖赏信号反馈 r_{t+1}，环境以一定概率转移到新的状态 s_{t+1}，等待 agent 做出下一步决策。在与环境的交互过程中，agent 有两处不确定性，一处是在状态 s 处选择什么样的动作，用策略 $\pi(a\mid s)$ 表示 agent 的某个策略（即状态到动作的概率分布）；另一处则是环境本身产生的状态转移概率 $P_{ss'}^{a}$，强化学习的目标是找到一个最优策略 π^*，使得它在任意状态 s 和任意时间步骤 t 都能够获得最大的长期累计奖赏，即：

$$\pi^*=\operatorname{argmax}_{\pi}\mathbb{E}_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k}\mid s_t=s\right\}\tag{1}$$

其中，$\mathbb{E}_{\pi}$ 表示策略下的期望值，$\gamma\in[0,1)$ 为折扣率（discount rate），k 为后续时间周期，r_{k+t} 表示 agent

表 1　　**各博弈环境关键属性**

游戏/兵棋	状态空间	动作空间	决策数量	胜利条件	回报值设置	战争迷雾	观察信息	对手建模	想定设计
《Go》	中等	中等	中等	数子法/数目法	简单	无	简单	中等	固定
《星际争霸Ⅱ》	复杂	复杂	较多	单任务目标	中等	有	中等	中等	变化较小
《Dota 2》	复杂	复杂	较多	单任务目标	中等	有	中等	中等	固定
《CMANO》	非常复杂	非常复杂	巨大	多任务目标	复杂	有	复杂	复杂	变化较大
《智戎•未来指挥官》	非常复杂	非常复杂	巨大	多任务目标/积分	复杂	有	复杂	复杂	变化较大
《王者荣耀》	复杂	复杂	较多	单任务目标	中等	有	中等	中等	固定
《战争游戏：红龙》	非常复杂	非常复杂	巨大	多任务目标	复杂	有	复杂	复杂	变化较大
《MaCA》	中等	中等	中等	积分	简单	有	中等	中等	固定

在时间周期$(t+k)$上获得的即时奖赏。

强化学习主要通过寻找最优状态值函数$V^*(s)$或最优状态动作值函数$Q^*(s,a)$来学习最优策略π^*。其中$V^*(s)$和$Q^*(s,a)$计算式如式（2）、式（3）所示：

$$V^*(s)=\max_\pi \mathbb{E}_\pi\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k}|\ s_t=s\right\} \tag{2}$$

$$Q^*(s,a)=\max_\pi \mathbb{E}_\pi\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k}|\ s_t=s,a_t=a\right\} \tag{3}$$

2.2　强化学习主流算法

2.2.1　基于值函数的强化学习

强化学习早期利用 Q-learning 算法来建立游戏 AI，通过预先设计每步动作可以获得的回报值来采取动作。Q-learning 最大的局限是需要提前设计好所有执行动作的回报值，它用一张 Q 表来保存所有的 Q 值，当动作空间巨大时，该算法难以适应。因此，Q-learning 算法只能在比较简单的环境中建模使用，如在简单的迷宫问题中，让 agent 通过 Q-learning 算法自动寻找出口。

DeepMind 在 2015 年第一次利用 DQN（deep Q network）算法在 Atari 游戏环境中实现了高水平的智能 AI，该 AI 综合评定达到了人类专业玩家的水平[24]。这也使得 DQN 算法成为强化学习的经典算法。DQN 算法通过神经网络拟合 Q 值，通过训练不断调整神经网络中的权重，获得精准的预测 Q 值，并通过最大的 Q 值进行动作选择。DQN 算法有效地解决了 Q-learning 算法中存储的 Q 值有限的问题，可以解决大量的离散动作估值问题，并且 DQN 算法主要使用经验回放机制（experience replay），即将每次和环境交互得到的奖励与状态更新情况都保存起来，用于后面的 Q 值更新，从而明显增强了算法的适应性。DQN 由于对价值函数做了近似表示，因此强化学习算法有了解决大规模强化学习问题的能力。但是 DQN 算法主要被应用于离散的动作空间，且 DQN 算法的训练不一定能保证 Q 值网络收敛，这就会导致在状态比较复杂的情况下，训练出的模型效果很差。在 DQN 算法的基础上，衍生出了一系列新的改进 DQN 算法，如 DDQN（double DQN）算法[51]、优先级经验回放 DQN（prioritized experience replay DQN）算法[52]、竞争构架 Q 网络（dueling DQN）算法[53]等。这些算法主要是在改进 Q 网络过拟合、改进经验回放中的采样机制、改进目标 Q 值计算等方面提升传统 DQN 算法网络的性能。总体来说，DQN 系列强化学习算法都属于基于值函数的强化学习算法类型。基于值函数的强化学习算法主要存在 3 点不足：对连续动作的处理能力不足、对受限状态下的问题处理能力不足、无法解决随机策略问题。由于这些原因，基于值函数的强化学习方法不能适用所有的场景，因此需要新的方法解决上述问题，例如基于策略的强化学习方法。

2.2.2　基于策略的强化学习

在基于值函数的强化学习方法中，主要是对价值函数进行了近似表示，引入了一个动作价值函数q，这个函数由参数w描述，以状态s与动作a为输入，计算后得到近似的动作价值，即式（4）：

$$\hat{q}(s,a,w)\approx q_\pi(s,a) \tag{4}$$

在基于策略的强化学习方法中，主要采用类似的思路，只不过主要对策略进行近似表示。此时，策略可以被描述为一个包含参数θ的函数，θ主要为神经网络中的权重，即式（5）：

$$\pi_\theta(s,a)=P(a|s,\theta)\approx\pi(a|s) \tag{5}$$

在基于策略的强化学习方法中，比较经典的就是理查德·萨顿（Richard S. Sutton）在 2000 年提出的 AC（actor-critic）框架强化学习算法。AC 包括两部分：演员（actor）和评价者（critic）。其中 actor 使用策略函数负责生成动作（action），通过动作与环境进行交互。而 critic 使用第 2.2.1 节提到的价值函数来评估 actor 的表现，并指导 actor 下一阶段的动作。总体来说，critic 通过 Q 网络计算状态的最优价值V_t，而 actor 利用V_t迭代更新策略函数的参数，进而选择动作，并得到反馈和新的状态，critic 使用反馈和新的状态更新 Q 网络参数w，在后面 critic 会使用新的网络参数w帮助 actor 计算状态的最优价值V_t。

2016 年 DeepMind 在国际机器学习大会（International Conference on Machine Learning）提出了 A3C 算法[54]。之前的 DQN 算法为了方便收敛使用了经验回放的技巧；AC 也可以使用经验回放的技巧。A3C 更进一步，还克服了一些经验回放的问题，主要采取随机性策略[55]。这里确定性策略和随机性策略是相对而言的，对于某一些动作集合来说，它可能是连续值，或者非常高维的离散值，这

导致动作的空间维度极大。如果使用随机性策略，即像DQN算法一样研究它所有的可能动作的概率，并计算各个可能动作的价值，那需要的样本量是非常大的。于是 DeepMind 就想出使用确定性策略来简化这个问题[56]。作为深度确定性策略梯度（deep deterministic policy gradient）、critic 目标网络和深度双 Q 网络（double DQN ）的当前 Q 网络，目标 Q 网络的功能定位基本类似，但是 DDPG 有自己的 actor 策略网络，因此不需要贪婪法这样的选择方法，这部分 DDQN 的功能到了 DDPG 可以在 actor 当前网络完成。而对经验回放池中采样的下一状态 s' 使用贪婪法选择动作 a'，这部分工作的作用是估计目标 Q 值，因此可以放到 actor 目标网络完成。

此外，actor 当前网络也会基于 critic 目标网络计算出的目标 Q 值进行网络参数的更新，并定期将网络参数复制到 actor 目标网络。DDPG 参考了 DDQN 的算法思想，通过双网络和经验回放，以及一些其他的优化，比较好地解决了 AC 难收敛的问题。因此在实际产品中尤其是与自动化相关的产品中使用得比较多，是一个比较成熟的 AC 算法。2017 年，Open AI 在神经信息处理系统大会（Conference and Workshop on Neural Information Processing Systems）上又提出了改进的多 agent 深度确定性策略梯度（multi-agent deep deterministic policy gradient）算法[57]，把强化学习算法进一步推广应用到多 agent 环境。在 AC 框架下，比较经典的算法还有近端策略优化（proximal policy optimization）算法[58]、柔性演员−评论家（soft actor-critic）算法[59]、双延迟深度确定性策略梯度（twin delayed deep deterministic policy gradient）算法[60]等，这些算法也都是在样本提取效率、探索能力增强方面进一步改进优化 AC 框架的。

2.3 深度学习结合强化学习

在现有策略对抗游戏中利用深度学习技术结合强化学习来实现游戏AI已成为主流研究方向[52]。其主要思路为在游戏对抗过程中利用图像特征的卷积提取技术。如在《觉悟 AI》中，图像特征的提取采取了分层的思想，在主视野和小地图中，对不同种类的要素进行提取并合并到一个层中，最终每层都提取到一类关键属性节点信息，形成“英雄”“野怪”“小兵”位置矩阵[6]。最终将多尺度特征的信息融合形成全局态势特征信息，这一工作同样应用在 AlphaStar 中。对于作战推演来说，态势理解一直是研究的难点，那么考虑利用深度学习技术来实现态势图像特征的提取，进而最终输出态势图的关键信息将是解决方案之一。此外，笔者所在团队也尝试利用深度学习技术对态势信息进行卷积提取，然后将提取信息与语义模型结合，生成当前态势的直观文本语义。而在前端利用强化学习进行实体单元控制，这样就可以将强化学习、深度学习、自然语言处理融合，在推演过程中实时生成方便人类理解的智能决策文本语义信息，这一工作对于实现推演系统中的人机融合具有积极意义。

2.4 分层强化学习

智能博弈对抗的建模过程面临两个难题，一个是动作空间庞大，另一个是奖励稀疏问题。面对这两个问题，有研究人员提出了分层强化学习的解决思路。该思路的核心是对动作进行分层，将低层级（low-level）动作组成高层级（high-level）动作，这样搜索空间就会被减小[52]。同时基于分层的思想，在一个预训练的环境中学习有用的技能，这些技能是通用的，与要解决的对抗任务的关系不太紧密。学习一个高层的控制策略能够使 agent 根据状态调用技能，并且该方法能够很好地解决探索效率较低的问题，该方法已在一系列稀疏奖励的任务中表现出色[61-62]。《觉悟 AI》同样设计了分层强化学习的动作标签来控制“英雄”的微观操作。具体来说，每个标签由两个层级（或子标签）组成，它们表示一级和二级操作。第一个动作，即一级动作，表示要采取的动作，包括移动、普通攻击、一技能、二技能、三技能、回血、回城等。第二个是二级动作，它告诉玩家如何根据动作类型具体地执行动作。例如，如果第一个层级是移动动作，那么第二个层级就是选择一个二维坐标来选择移动的方向；当第一个层级为普通攻击时，第二个层级将成为选择攻击目标；如果第一个层级是一技能（或二技能、三技能），那么第二个层级将针对不同技能选择释放技能的类型、目标和区域。这对于作战推演中不同算子如何执行动作也具有参考价值，每一个类型的算子同样存在不同的动作，例如坦克可以选择直瞄射击、间瞄射击、移动方向等，实际作战推演不同装备同样具有众多复杂的动作，通过这样的特征和标签设计，可以将人工智能建模任务作为一个层次化的多类分类问题来完成。具体来说，一个深层

次的神经网络模型被训练以预测在给定的情境下要采取的行动。作战推演也可以参考层次化的动作标签来不断细化动作执行过程，进而训练解决复杂的动作执行难题。在作战推演中完全可以借鉴这种思路设计适用于作战场景的分层强化学习框架。南京大学的研究人员利用分层强化学习建立宏观策略模型和微观策略模型，根据具体的态势评估宏观策略模型，然后利用宏函数批量绑定选择微观动作，这样可以在不同的局势下选择对应的一系列动作，进而实现了分层强化学习在《星际争霸》环境中的应用[63]。分层强化学习比较通用的框架是两层，顶层策略被称为元控制器（meta-controller），负责生成总体宏观目标，底层策略被称为控制器（controller），负责完成给定的子目标，这种机制本质也对应作战推演中的战略、战役、战术 3 个层次，不同层次关注的作战目标各有不同，但又互相关联。其他相关改进是学者在奖赏函数设置、增加分层结构、保持分层同步、提高采样效率等方面改进分层强化学习[64]。

2.5　多 agent 强化学习

在游戏博弈对抗过程中必然需要考虑多 agent 建模，而在作战推演中利用多 agent 技术实现不同作战单元的协同合作也是博弈智能研究的重点之一。在这方面 OpenAI 和 AlphaStar 在多 agent 深度强化学习方面使用了不同的技术思路。OpenAI 使用的是分布异构的多 agent 建模思路，每一个 agent 都有一个相同的训练神经网络，但是没有全局控制网络[37,47]。AlphaStar 则是使用了一个集中的控制网络对不同的单元进行控制。还有一种思路是对于每一个 agent，都建立属于各自的神经网络进行训练。第三种思路是最理想的，但是训练过程复杂，也难以适用于大规模的推演过程[43]。对于实际作战推演来说，除了要考虑多 agent 建模方法，还需要让每个 agent 具有柔性加入的能力，在对抗过程中可以按照需要随时加入所需要的作战单元，而不需要每次加入作战单元后，再重新训练一遍网络。基于此考虑，让每一个 agent 具有自己独立的神经网络将是更好的选择。

2.6　LSTM 技术结合深度强化学习

《觉悟 AI》在设计过程中利用深度学习不断提取游戏界面的态势信息。利用深度学习虽然可以把一个对抗界面的所有特征提取出来，但是提取的是静态的某一帧的界面信息，并没有把时间步之间的信息关联起来。时间步一般指一帧，也可以指多帧，其关键是将历史的帧信息和现在的信息关联起来。基于此，需要引入长短期记忆（long short-term memory，LSTM）网络。让 LSTM 一次接收多个时间步信息来学习这些时间步之间的关联信息，从而让 LSTM 帮助“英雄”学习技能组合，并选择“英雄”应该关注的主视野和小地图的具体方面，进而综合输出合理的动作，也通过 LSTM 关联历史数据来训练强化学习的神经网络模型[65]。在实际作战推演过程中同样需要考虑这种情况，防止出现训练的 AI 为了某个战术目标而忽视了整体战略目标。

2.7　多属性决策结合强化学习

强化学习的回报值往往根据专家经验手工设置，但是这种手工设置的回报值的收敛性及智能性均难以保证，并且长时间训练才能评估设置的回报值的效果。因此可以考虑结合推演数据，结合多属性决策方法进行客观分析，总结提炼出合适的回报值。首先，从推演环境中获取各关键属性数据，如在陆战对抗环境提取作战单元位置、高程、类型、射程属性、打击属性、装甲属性等。以这些属性数据为基础，计算出对应的评估指标，如目标距离威胁、目标攻击威胁、目标速度威胁等，通过熵权法计算相应权重，并最终结合多属性方法对敌方威胁度进行排序，将其和强化学习的回报值函数进行关联，进而设置出更加客观合理的回报值函数，这样有利于提高强化学习训练的智能性，并有利于加快收敛。

3　其他可用智能决策技术

在智能博弈领域，国际上的 Atari、AlphaGo、AlphaStar、OpenAI 都取得了显著的成果，国内《觉悟 AI》《墨子•未来指挥官系统》AI、《CASIA-先知 1.0》也都取得了突破性的进展。这些工作主要以深度强化学习技术为主，但均搭配使用了其他相关的人工智能技术。总体来说，单纯地利用深度强化学习技术并不能很有效地实现智能 AI，有必要在训练过程中结合其他技术提高 AI 性能。同时，如果想要实现特别突出的 AI 智能，那么将在训练的过程中花费大量的成本。AlphaStar 的训练过程持续了 10 个月，使用了 51 000 个 CPU。该工作同时有 30 个博士生参与，花费数百万美元的成本，才在游戏智能博弈领域实现了超过职业选手水平的 AI 智能[43]。

3.1 进化算法

借鉴生物进化论，遗传算法将要解决的问题模拟成一个生化进化的过程，通过复制、交叉、突变等操作产生下一代的解，并逐步淘汰适应度函数值低的解，增加适应度函数值高的解。这样进化 N 代后就有可能进化出适应度函数值很高的个体。

在 1993 年就有人尝试用遗传算法训练神经网络，但是当时的计算机算力不足，导致这个方向并没有引起过多关注[66-68]。随着深度强化学习技术的火热发展以及算力的显著提高，部分学者和机构又开始关注这一结合点。OpenAI 在 2017 年尝试直接利用进化算法来代替强化学习技术，在 MuJoCo（multi-joint dynamics with contact）和 Atari 上取得了一定的效果。但是这一工作的前提是需要大量的 CPU 进行大规模训练，且实验环境比较简单[69]。优步（Uber）AI 在 2017 年尝试将基于种群的遗传算法和深度神经网络结合，利用进化策略而不是梯度策略来更新权重参数，取得的算法性能在一定程度上优于 A3C、DQN 算法[70]。除了在优化网络参数方面进行结合，将进化策略和多 agent 强化学习结合也是一个有意义的方向。DeepMind 就是在 AlphaStar 中利用了联盟赛制，在训练出的 agent 中不断优化筛选出更加优秀的 agent，从而不断演化，最终训练出超过职业选手水平的游戏 AI[48]。

总体来说，相较于强化学习算法，单纯利用遗传算法有明显的缺陷，遗传算法采样效率过低，并且不可以按照梯度优化的方式进行参数调整。遗传算法在实际推演中可能需要每一局结束才可以更新一个策略或者优化一个动作，而不能像强化学习算法在推演中每一步都进行一定的更新。但是，遗传算法的优势是适合在大规模的空间中进行探索，寻找全局最优解。而强化学习算法随着梯度下降进行优化，很容易寻找到局部最优解，而不是全局最优解。因此，如果找到合适的结合角度，遗传算法和强化学习两者结合会有一定的实际价值。

3.2 决策树

决策树方法是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，并判断其可行性的决策分析方法，是直观运用概率分析的一种图解法[71]。其本身是一种树形结构，其中每个内部节点表示一个行为判断，每个分支代表一个判断结果的输出，每个叶节点代表一种分类结果。

在作战推演建模的早期研究中，决策树是一种非常重要且常用的建模方法[72]，其易于构建作战实体的行为规则，有利于分析基于决策树的作战实体行为模型[73]，这在作战推演的初期是一种快速建立对手模型的高效办法，并且基于决策树的作战 AI 也具有一定的初步智能性。在现在的游戏智能博弈对抗过程中，虽然基于决策树的研究总体比较少，但是衍生出了一些重要的算法，如南京大学 Zhou Z H 等人[74]提出的深度森林算法就是在决策树的基础上拓展而来的。这些工作都为后面的智能博弈领域的研究提供了重要的理论基础。

3.3 基于规则

基于规则的 AI（简称规则 AI）主要是结合博弈对抗环境的领域知识，构建基于专家经验知识的规则 AI。这类规则 AI 主要以高水平玩家的经验知识为基础，对领域知识进行程序化，进而形成具有一定智能性的推演 AI。近年，在国内的智能博弈比赛中，参赛团队还是普遍以规则 AI 为基础，对规则 AI 进行改良和设计进而参赛[3]。在 2020 年的智能兵棋比赛中，大部分团队及基准 AI 还是以规则 AI 为主流。当然，单纯基于知识的规则也存在各种局限，如智能性普遍较低、通用性较差等。但是，规则 AI 的好处就是便于分析设计，可以快速构建具有一定智能性的博弈对抗 AI 环境。规则 AI 可以作为对手模型进行构建，让强化学习 AI 与规则 AI 进行对抗，初步验证强化学习的智能性。同时，通过规则驱动结合强化学习的智能 AI 构建，也是当前国内智能兵棋的研究热点，利用高水平玩家快速构建基于规则的 AI，让 agent 快速学习有效动作并存入模型中，方便神经网络直接提取有效经验，进而实现强化学习的快速收敛，加快学习进程。国内也有研究尝试利用知识驱动结合数据驱动，通过知识牵引 AI 的整体策略，以数据驱动 AI 的具体动作，设计出基于知识牵引与数据驱动的兵棋 AI 框架[18]。

3.4 势能统计

国内研究人员借鉴物理学中的势能理论与方法，对指挥决策人员与战场要素间作用关系及其发展趋势进行量化分析和形式化表达，引导智能决策实体进行行动策略选择。并且从势能角度分析了作战指挥决策机理，尝试利用基于变权的动态势能模型和基于统计分析的静态势能模型，构建了基于综合势能的作战行动序列生成方法。并尝试在智能兵棋领域进行实验，验证了该算法优于多数规则及知

识 AI[75]。实际上，该方法是利用离线和在线统计数据综合分析出智能兵棋推演 AI 的。可以尝试将该方法与强化学习结合，弥补强化学习开始阶段训练收敛速度过慢的缺陷，并在强化学习算法执行过程中，结合综合势能进行动作校正，从而生成更加智能化的作战行动序列。

3.5　随机森林

在机器学习中，随机森林是一个包含多个决策树的分类器，并且输出的类别是由个别树输出的类别的众数而定的。Breiman L[76]推论出随机森林算法。随机森林的随机性主要体现在可以从原始数据中采取有放回的抽样构造子数据集，子数据集的数据量和原始数据集是相同的。随机森林可以随机选取待选特征，随机森林中的子树的每一个分裂过程并未用到所有的待选特征，而是从所有的待选特征中随机选取一定的特征，再在随机选取的特征中选取最优的特征。随机森林的优点是实现相对简单，训练速度较快，可以并行实现；相比单一的决策树，能学习到特征之间的相互影响，并且不容易过拟合；对于高维数据来说，不需要做特征选择，明显提高了训练效率。

在智能博弈对抗领域，与随机森林相关的研究其实较少，但是其在一定程度上可以作为训练数据的有效手段，进而弥补一些强化学习算法训练效率较低的缺陷。已有学者在德州扑克的博弈对抗环境中将策略空间设计为一种快速且高效的决策树，进而有利于使用多种方法来学习这种启发式的方法[77]。

3.6　人件技术

真实、完整地刻画人的直接参与特点，并且对人的服务角色进行统一管理和调度是构建人机融合系统的重要前提，人件技术能够更好地把人真正融入人机交互系统中，使该系统真正体现以人为本、强调智能的新型交互系统特点[78-79]。在智能博弈对抗环境中，人件技术主要是在专家经验知识中进行考虑，主要利用高水平玩家的数据进行监督学习，方便快速高效地训练出高水平的深度强化学习 AI。人件技术的核心是在训练过程中融入人的行为偏好，通过人类行为决策数据进行训练，训练出一个初步的模型。而强化学习算法可以直接从初步的模型中提取相关数据，进而能训练出更具有智能性的 AI。DeepMind 对 AlphaStar 做了一组关于专家经验的消融实验，该消融实验证明 AlphaStar 在复杂环境中，单纯依靠强化学习进行训练很难获得效果。同时该实验证明在仅有监督学习技术的支持下，AlphaStar 可以达到较好的效果。在充分利用人类数据后，AlphaStar 的性能可以再次提高 60%。

3.7　统计前向规划算法

统计前向规划算法使用仿真模型（也称为前向模型）自适应地搜索有效的动作序列，此类算法提供了一种简单通用的方法，为各种游戏提供快速自适应的 AI 控制。常见的经典模型为 MCTS 算法，MCTS 算法最重要的优点是不需要领域特定知识，可以在不了解游戏规则的情况下应用，这使得它很容易适用于任何可以使用树进行建模的领域。像 Go[25]这样的游戏，分支因子明显具有高数量级特征，而有用的启发式又很难形成，这类问题就需要使用 MCTS 算法解决[80]。尽管 MCTS 算法在大范围的博弈中提供了更强的决策能力，但将其应用在作战推演领域仍存在很多挑战和瓶颈。在作战推演领域，当需要搜索的图的分支因子和深度被限定，作战推演非常耗费 CPU、GPU 资源时，MCTS 算法是否仍然是指导作战推演的最佳方法是一个有待研究的问题。

3.8　小地图设置技术

在多个智能博弈对抗游戏中普遍存在一个小地图，用来辅助玩家快速了解整体态势。AlphaStar 利用 ResNet 在小地图中进行特征提取，获得对抗博弈中的关键属性信息，最终形成一个离散的单元特征图。AlphaStar 正是通过小地图+单位列表+标量信息（资源信息）汇总输出各种智能决策给出的执行方案。在实际作战推演中，也需要考虑针对某个战场的全局地图信息，指挥员可能关注某个局部作战场景，同时也应该考虑全局的作战信息的获取[27]。因此，在作战推演中智能决策 AI 的训练也需要设计小地图机制，来辅助深度强化学习智能 AI 进行训练。

4　作战推演技术难点及技术解决方案

4.1　冷启动问题

在对强化学习的训练过程中，研究人员总是会遇到训练过程时间长、难以收敛的问题，这种情况通常被称为冷启动问题。针对这个问题，现有研究人员提出了多种解决方案，比较有效的解决方案是利用专家的领域知识预先设计固定的先验知识，利用先验知识进行智能博弈训练，进而在强化学习的

经验存储中得到高水平的训练数据。在强化学习的后期训练中直接利用这些先验知识对抗出来的经验数据进行模型训练，从而可以有效缩小探索空间和动作空间，进而保证强化学习可以快速训练出高水平的 AI，避免了前期盲目探索的情况。在实际作战推演过程中，也可以考虑使用高水平指挥员的先验知识，提前进行形式化存储，进而在强化学习训练过程中导入先验知识，加快训练结果的收敛，得到较高水平的智能 AI。

4.2 过拟合问题

在智能博弈对抗过程中经常会出现训练一定阶段后陷入局部最优结果的情况。表现为在智能兵棋比赛中，经过长时间训练后，强化学习训练出的结果是控制算子进行固定的线路和射击套路，这种情况可被称为过拟合现象。为了避免这种情况的出现，应该在算法设计中加入随机可能性，在一定比例的动作选择概率下随机探索，而不是完全按照强化学习算法给出的结果进行执行。其次，按照贝尔曼方程，应该在奖励函数设计过程中，考虑当前影响和未来影响的可变比重，即回报函数设计包括一定的可变性，而不是固定不变的。当然也可以利用强大的计算力，生成大量新的对手，从不同方面与需要训练的 agent 进行对抗，从而避免因为固定对手而导致的过拟合现象。

4.3 想定适应性问题

智能博弈的 AI 建模普遍存在适应性不高的问题，有部分研究人员开发的 AI 是针对某个固定想定开发的，导致更换博弈想定后 AI 性能大幅下降。考虑到大部分数据或任务是存在相关性的，通过迁移学习可以将已经学到的模型参数通过某种方式分享给新模型，从而加快优化模型效率。中国科学院自动化研究所的研究人员引入了课程迁移学习，将强化学习模型扩展到各种不同博弈场景，并且提升了采样效率[81]。DeepMind 在 AlphaZero 中使用同样的算法设置、网络架构和超参数，得到了一种适用于围棋、国际象棋和将棋的通用算法，并战胜了基于其他技术的棋类游戏 AI[82]。《觉悟 AI》引入了课程学习方法，将训练至符合要求的参数迁移至同一个神经网络再次训练、迭代、修正以提高效率，使《觉悟 AI》模型能熟练掌握 40 多个“英雄”[6,36]。在作战推演中，更需要这种适用性强的通用 AI 算法，不需要在更换作战想定后重新训练模型，也只有这样才可以更加适应实时性要求极高的作战场景。

4.4 智能蓝方建模

对手建模指在两个 agent 博弈的环境中，为了获得更高的收益，需要对对手的策略进行建模，利用模型（隐式）推断其所采取的策略来辅助决策。智能蓝方建模主要是在具有战争迷雾的情况下，对对手进行建模，并预测对手的未来动作。其前提通常是博弈环境存在战争迷雾，我方在无法获取准确的对手信息的情况下，针对对方进行预测评估。在对抗过程中，一种假设是对手是完全理性的，对对手（队友）进行建模是为了寻找博弈中的纳什均衡策略。为了解决这一难点问题，阿尔伯塔大学的研究人员提出了反事实遗憾最小化（counterfactual regret minimization，CFR）技术，该技术不再需要一次性推理一棵完整的博弈树，而是允许从博弈的当前状态使用启发式搜索。另外，对手建模可分为隐式建模和显式建模。通常隐式建模直接将对手信息作为自身博弈模型的一部分来处理对手信息缺失的问题，通过最大化 agent 期望回报的方式将对手的决策行为隐式引进自身模型，构成隐式建模方法。显式建模则直接根据观测到的对手历史行为数据进行推理优化，通过模型拟合对手行为策略，掌握对手意图，降低对手信息缺失带来的影响[83]。总体来说，对手建模技术是智能博弈对抗是否有效的关键，只有建立一个可以高效预估对手行为的模型，才能保证智能博弈 AI 的有效性。

4.5 路径规划问题

路径规划作为智能博弈中的重要组成部分，其主要任务是根据不同的想定，针对每个单元在起始点和终止点之间快速规划一条由多个路径点依次连接而成的最优路径[84]。在智能博弈的背景下，最优路径的含义不仅仅是两点之间的距离最短，而是综合考虑博弈态势、资源情况和综合威胁后的最佳路径。但是，已有的路径规划算法主要以 A-Star 算法、Dijkstra 算法、D*算法、LPA*算法、D* lite 算法等为典型代表，在物流运输、无人驾驶、航空航天等领域都取得了显著成效。同时也有学者提出其他的路径规划算法，如基于神经网络和人工势场的协同博弈路径规划方法[85]等，但是在智能博弈的环境下，需要考虑的问题更加复杂，需要进一步对这些算法进行改进优化。

5　作战推演未来发展建议

5.1　智能作战推演通用框架

现有的游戏平台中也有比较成熟的AI开发通用框架（如 pysc2[86-87]），但是相比成熟的作战推演通用框架还有较大差距。智能作战推演系统可以设计一个适用于复杂环境的通用框架，该框架包括作战推演算子、地图、规则、想定。同时最关键的是设计通用的算法接口，这些接口可以方便智能博弈算法的设计与实现，如环境加载接口、环境重置接口、环境渲染接口、动作随机选择接口、执行动作接口等。同时，也可以提前设计智能作战推演的基本功能框架，包括地图编辑模块、想定编辑模块、算子管理模块、规则编辑模块、推演设置模块、数据分析模块、系统配置模块。其中最核心的是推演设置模块可以自由选择每局推演使用的智能算法，从而实现智能算法设计和作战推演环境开发的解耦，这样才可以保证智能作战推演的灵活适应性。通用框架中另一个重要的因素是可以提供AI 使用的工具，例如对于深度学习的分层态势显示，可以直观地提供一个通用接口进行展现，方便指挥人员快速调用该接口实现智能算法的辅助决策功能。

5.2　智能战略、战役、战术决策方案制定

智能作战推演必然面对的问题是选择在战略、战役还是战术场景下应用。现阶段主要还是在战术层面进行智能算法的研究，包括国内的某智能兵棋推演大赛，各种想定只有算子数量种类的差别，但本质上都还属于战术智能决策。《墨子•未来指挥官系统》中的对抗想定更接近于战役层面的智能决策方案，现阶段对战略层面的智能决策的研究还比较少。其原因就在于面临的想定越宏观，智能决策的技术挑战越大，包括动作空间、状态空间的变化以及现阶段 agent 之间的协同交互还并没有很成熟。因此，当前考虑战术层面的智能决策更容易。如果要解决战略层面的智能决策，必然需要研究各 agent 之间的协同机制，还要考虑作战的后勤支持机制。然而当前尚未有游戏、作战推演在智能推演中考虑后勤机制的影响。另外，战术、战役、战略层面的方案制定技术思路也并不相同，有的研究以各 agent 独自训练、交互，进而涌现出智能决策方案。这一技术思路更加逼近真实场景，但是算力要求和技术实现难度都较高。另一思路是建立统一的宏观 agent 模型，利用宏观 agent 控制所有算子进行推演，这一技术思路实现比较简单，所需的算力也较低，可以考虑将其作为初期实现的路径之一。

5.3　人机融合的智能作战推演模式建立

智能作战推演的未来趋势主要分为人不在环和人在环两种类型。人不在环主要类似于 AlphaStar、OpenAI 的游戏智能，通过预先训练完成 agent，完全由训练好的 agent 自主进行博弈对抗，左右互搏，实现方案的预演和推测。人在环的模式又分为两种，一种是实现人机对抗，国内已有这方面的比赛，通过开发训练好的智能算法 agent 与高水平指挥人员进行对抗，探测发现自身指挥问题并不断提高，可用于指挥人员训练场景。另一种人在环更加困难，即 agent 可以响应人的指令，并完成低层次的规划任务。主要还是由指挥人员进行整体战略宏观判断，并通过指令交互部署 agent 完成低层次任务，最后总体实现战略目标。同时，也需要对人机融合模式的框架进行探索，如将人类领域知识引入智能算法中，帮助智能算法更高效地实现智能作战推演。

5.4　开放性的仿真实验平台建立

随着智能博弈的兴起，国内外高校、研究所、企业都已逐渐开发完成各种类型的智能博弈平台。但是不同平台之间并不互通，相互独立，形成了各个平台的信息孤岛，在智能算法开发的过程中，面对新的平台就需要研究人员重新学习适应新的平台接口和架构，这浪费了研究人员大部分精力。另外，智能博弈的强化学习接口以及其他算法虽然在不同平台体现不同，但本质都一样，很有必要构建一个通用一体化智能博弈平台框架，减少不断重新开发、学习的过程，提高智能博弈平台的研究效率也势在必行。

6　结束语

本文构建并梳理了智能作战推演的国内外发展概况，详细介绍了智能作战推演所需要的各项技术难点及国内外进展，同时借鉴游戏 AI 领域的发展现状并将其与智能作战推演所需要的技术需求进行对比，分析现有技术还需要改进和优化的方向，也对智能作战推演的各个研究方向进行了归纳。可以为从事游戏 AI、智能兵棋、智能作战推演等智能博弈领域的研究人员提供一定的思路。当前智能博弈的研究思路还主要以深度强化学习为基础，但绝不仅仅是深度强化学习技术，各种传统的

智能算法和新的机器学习算法都可以作为智能博弈领域补充完善的技术力量。本文以作战推演技术需求为牵引，结合国内外游戏 AI、兵棋推演相关实际发展情况，总结出虽然智能博弈依然还有很多难题需要解决，现有技术实现程度相比于实际应用还有较大差距，但智能博弈这一研究方向是未来智能决策研究发展的必由之路，相信最终会在各个相关领域得以实现。

参考文献：

[1] 沈宇, 韩金朋, 李灵犀, 等. 游戏智能中的 AI: 从多角色博弈到平行博弈[J]. 智能科学与技术学报, 2020, 2(3): 205-213.
SHEN Y, HAN J P, LI L X, et al. AI in game intelligence—from multi-role game to parallel game[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(3): 205-213.

[2] 胡晓峰, 贺筱媛, 陶九阳. AlphaGo 的突破与兵棋推演的挑战[J]. 科技导报, 2017, 35(21): 49-60.
HU X F, HE X Y, TAO J Y. AlphaGo's breakthrough and challenges of wargaming[J]. Science & Technology Review, 2017, 35(21): 49-60.

[3] 叶利民, 龚立, 刘忠. 兵棋推演系统设计与建模研究[J]. 计算机与数字工程, 2011, 39(12): 58-61.
YE L M, GONG L, LIU Z. Research and modeling of a rehearsal system of naval battle[J]. Computer & Digital Engineering, 2011, 39(12): 58-61.

[4] 谭鑫. 基于规则的计算机兵棋系统技术研究[D]. 长沙: 国防科学技术大学, 2010.
TAN X. Research on rule-based computer wargame system technology[D]. Changsha: National University of Defense Technology, 2010.

[5] 胡晓峰, 齐大伟. 智能决策问题探讨——从游戏博弈到作战指挥, 距离还有多远[J]. 指挥与控制学报, 2020, 6(4): 356-363.
HU X F, QI D W. On problems of intelligent decision-making—how far is it from game-playing to operational command[J]. Journal of Command and Control, 2020, 6(4): 356-363.

[6] YE D H, CHEN G B, ZHAO P L, et al. Supervised learning achieves human-level performance in MOBA games: a case study of honor of kings[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020: 1-11.

[7] FU H T, TANG H Y, HAO J Y, et al. Deep multi-agent reinforcement learning with discrete-continuous hybrid action spaces[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2019.

[8] WANG X J, SONG J X, QI P H, et al. SCC: an efficient deep reinforcement learning agent mastering the game of StarCraft II[J]. arXiv preprint, 2020, arXiv:2012.13169.

[9] 周超, 胡晓峰, 郑书奎, 等. 战略战役兵棋演习系统兵力聚合问题研究[J]. 指挥与控制学报, 2017, 3(1): 19-26.
ZHOU C, HU X F, ZHENG S K, et al. Force integration in strategic and operational war-game maneuver system[J]. Journal of Command and Control, 2017, 3(1): 19-26.

[10] 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 technologies of human-computer gaming[J]. Scientia Sinica (Informationis), 2020, 50(4): 540-550.

[11] LIU X, ZHAO M J, DAI S, et al. Tactical intention recognition in wargame[C]//Proceedings of 2021 IEEE 6th International Conference on Computer and Communication Systems. Piscataway: IEEE Press, 2021: 429-434.

[12] SUN Y X, YUAN B, ZHANG T, et al. Research and implementation of intelligent decision based on a priori knowledge and DQN algorithms in wargame environment[J]. Electronics, 2020, 9(10): 1668.

[13] 陈希亮, 李清伟, 孙彧. 基于博弈对抗的空战智能决策关键技术[J].指挥信息系统与技术, 2021, 12(2): 1-6.
CHEN X L, LI Q W, SUN Y. Key technologies for air combat intelligent decision based on game confrontation[J]. Command Information System and Technology, 2021, 12(2): 1-6.

[14] 孙彧, 李清伟, 徐志雄, 等. 基于多智能体深度强化学习的空战博弈对抗策略训练模型[J]. 指挥信息系统与技术, 2021, 12(2): 16-20.
SUN Y, LI Q W, XU Z X, et al. Game confrontation strategy training model for air combat based on multi-agent deep reinforcement learning[J]. Command Information System and Technology, 2021, 12(2): 16-20.

[15] 瞿崇晓, 高翔, 夏少杰, 等. 一种基于深度强化学习的无监督智能作战推演系统: CN109636699A[P]. 2019.
QU C X, GAO X, XIA S J, et al. Unsupervised intelligent combat deduction system based on deep reinforcement learning: CN109636699A[P]. 2019.

[16] 张振, 黄炎焱, 张永亮, 等. 基于近端策略优化的作战实体博弈对抗算法[J]. 南京理工大学学报, 2021, 45(1): 77-83.
ZHANG Z, HUANG Y Y, ZHANG Y L, et al. Battle entity confrontation algorithm based on proximal policy optimization[J]. Journal of Nanjing University of Science and Technology, 2021, 45(1): 77-83.

[17] 李琛, 黄炎焱, 张永亮, 等. Actor-Critic 框架下的多智能体决策方法及其在兵棋上的应用[J]. 系统工程与电子技术, 2021, 43(3): 755-762.
LI C, HUANG Y Y, ZHANG Y L, et al. Multi-agent decision-making method based on Actor-Critic framework and its application in wargame[J]. Systems Engineering and Electronics, 2021, 43(3): 755-762.

[18] 程恺, 陈刚, 余晓晗, 等. 知识牵引与数据驱动的兵棋 AI 设计及关键技术[J]. 系统工程与电子技术, 2021, 43(10): 2911-2917.
CHENG K, CHEN G, YU X H, et al. Knowledge traction and data-driven wargame AI design and key technologies[J]. Systems Engineering and Electronics, 2021, 43(10): 2911-2917.

[19] 张可, 郝文宁, 余晓晗, 等. 基于遗传模糊系统的兵棋推演关键点推理方法[J]. 系统工程与电子技术, 2020, 42(10): 2303-2311.
ZHANG K, HAO W M, YU X H, et al. Wargame key point reasoning method based on genetic fuzzy system[J]. Systems Engineering and Electronics, 2020, 42(10): 2303-2311.

[20] 李航, 刘代金, 刘禹. 军事智能博弈对抗系统设计框架研究[J]. 火力与指挥控制, 2020, 45(9): 116-121.
LI H, LIU D J, LIU Y. Architecture design research of military intelligent wargame system[J]. Fire Control & Command Control, 2020, 45(9): 116-121.

[21] 施伟, 冯旸赫, 程光权, 等. 基于深度强化学习的多机协同空战方法研究[J]. 自动化学报, 2021, 47(7): 1610-1623.
SHI W, FENG Y H, CHENG G Q, et al. Research on multi-aircraft

cooperative air combat method based on deep reinforcement learning[J]. Acta Automatica Sinica, 2021, 47(7): 1610-1623.

[22] 徐佳乐, 张海东, 赵东海, 等. 基于卷积神经网络的陆战兵棋战术机动策略学习[J]. 系统仿真学报, 2021: 已录用.
XU J L, ZHANG H D, ZHAO D H, et al. Tactical maneuver strategy learning of wargame based on convolutional neural network[J]. Journal of System Simulation, 2021: acceped.

[23] WANG H N, LIU N, ZHANG Y Y, et al. Deep reinforcement learning: a survey[J]. Frontiers of Information Technology & Electronic Engineering, 2020, 21(12): 1726-1744.

[24] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[25] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[26] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[27] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature, 2019, 575(7782): 350-354.

[28] BERNER C, BROCKMAN G, CHAN B, et al. Dota 2 with large scale deep reinforcement learning[J]. arXiv preprint, 2019, arXiv:1912.06680.

[29] BROWN N, SANDHOLM T. Superhuman AI for multiplayer poker[J]. Science, 2019, 365(6456): 885-890.

[30] SCHRITTWIESER J, ANTONOGLOU I, HUBERT T, et al. Mastering Atari, Go, chess and shogi by planning with a learned model[J]. Nature, 2020, 588(7839): 604-609.

[31] PRICE M. What impact do VR controllers have on the traditional strategy game genre[D]. Huddersfield: University of Huddersfield, 2019.

[32] DAVID A S, JOHNSON M. Reinforcing deterrence on NATO's eastern flank: wargaming the defense of the baltics[R]. 2016.

[33] CANNON C T, GOERICKE S. Using convolution neural networks to develop robust combat behaviors through reinforcement learning[D]. CA: Naval Postgraduate School, 2021.

[34] 绵珊珊. 美俄人工智能军事应用发展分析[J]. 大数据, 2020, 6(4): 125-132.
XIAN S S. An analysis of the military application and development path of artificial intelligence in the United States and Russia[J]. Big Data Research, 2020, 6(4): 125-132.

[35] TARRAF D C, GILMORE J M, BOSTON S. An experiment in tactical wargaming with platforms enabled by artificial intelligence[R]. 2020.

[36] YE D H, LIU Z, SUN M F, et al. Mastering complex control in MOBA games with deep reinforcement learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 6672-6679.

[37] BROCKMAN G, CHEUNG V, PETTERSSON L, et al. OpenAI gym[J]. arXiv preprint, 2016, arXiv:1606.01540.

[38] ARULKUMARAN K, CULLY A, TOGELIUS J. Alphastar: an evolutionary computation perspective[C]//Proceedings of the Genetic and Evolutionary Computation Conference Companion. [S.l.:s.n.], 2019: 314-315.

[39] YE D H, CHEN G B, ZHANG W, et al. Towards playing full MOBA games with deep reinforcement learning[J]. arXiv preprint, 2020, arXiv:2011.12692.

[40] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[J]. arXiv preprint, 2013, arXiv:1312.5602.

[41] 张凯峰, 俞扬. 基于逆强化学习的示教学习方法综述[J]. 计算机研究与发展, 2019, 56(2): 254-261.
ZHANG K F, YU Y. Methodologies for imitation learning via inverse reinforcement learning: a review[J]. Journal of Computer Research and Development, 2019, 56(2): 254-261.

[42] 曹雷. 基于深度强化学习的智能博弈对抗关键技术[J]. 指挥信息系统与技术, 2019, 10(5): 1-7.
CAO L. Key technologies of intelligent game confrontation based on deep reinforcement learning[J]. Command Information System and Technology, 2019, 10(5): 1-7.

[43] RISI S, PREUSS M. Behind DeepMind's AlphaStar AI that reached grandmaster level in StarCraft II[J]. KI-KünstlicheIntelligenz, 2020, 34(1): 85-86.

[44] SILVER D, VENESS J. Monte-Carlo planning in large POMDPs[C]//Proceedings of the Advances in Neural Information Processing Systems 23. [S.l.:s.n.], 2010.

[45] GOODMAN J, LUCAS S. Does it matter how well I know what you're thinking? Opponent modelling in an RTS game[C]//Proceedings of 2020 IEEE Congress on Evolutionary Computation. Piscataway: IEEE Press, 2020: 1-8.

[46] JOHANSON M. Measuring the size of large no-limit poker games[J]. arXiv preprint, 2013, arXiv:1302.7008.

[47] DUGAS D, NIETO J, SIEGWART R, et al. Navrep: unsupervised representations for reinforcement learning of robot navigation in dynamic human environments[C]//Proceedings of 2021 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2021: 7829-7835.

[48] ONTANÓN S, SYNNAEVE G, URIARTE A, et al. A survey of real-time strategy game AI research and competition in StarCraft[J]. IEEE Transactions on Computational Intelligence and AI in games, 2013, 5(4): 293-311.

[49] FENNER S A, ROGERS J. Combinatorial game complexity: an introduction with poset games[J]. arXiv preprint, 2015, arXiv: 1505.07416.

[50] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[J]. IEEE Transactions on Neural Networks, 2005, 16(1): 285-286.

[51] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double q-learning[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Piscataway: IEEE Press, 2016.

[52] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay[J]. arXiv preprint, 2015, arXiv:1511.05952.

[53] WANG Z Y, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning[J]. arXiv preprint, 2015, arXiv:1511.06581.

[54] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[C]//Proceedings of the 33rd International Conference on Machine Learning. [S.l.:s.n.], 2016: 1928-1937.

[55] 刘朝阳, 穆朝絮, 孙长银. 深度强化学习算法与应用研究现状综述[J]. 智能科学与技术学报, 2020, 2(4): 314-326.
LIU Z Y, MU C X, SUN C Y. An overview on algorithms and applications of deep reinforcement learning[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(4): 314-326.

[56] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control

with deep reinforcement learning[J]. arXiv preprint, 2015, arXiv: 1509.02971.

[57] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C]//Proceedings of the Advances in Neural Information Processing Systems 30. [S.l.:s.n.], 2018.

[58] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv preprint, 2017, arXiv:1707.06347.

[59] 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. [S.l.:s.n.], 2018: 1861-1870.

[60] FUJIMOTO S, VAN HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[C]//Proceedings of the International Conference on Machine Learning. [S.l.:s.n.], 2018: 1587-1596.

[61] FLORENSA C, DUAN Y, ABBEEL P. Stochastic neural networks for hierarchical reinforcement learning[J]. arXiv preprint, 2017, arXiv: 1704.03012.

[62] RAFATI J, NOELLE D C. Learning representations in model-free hierarchical reinforcement learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 10009-10010.

[63] PANG Z J, LIU R Z, MENG Z Y, et al. On reinforcement learning for full-length game of StarCraft[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 4691-4698.

[64] LI S Y, WANG R, TANG M X, et al. Hierarchical reinforcement learning with advantage-based auxiliary rewards[J]. arXiv preprint, 2019, arXiv:1910.04450.

[65] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[66] YAO X. A review of evolutionary artificial neural networks[J]. International Journal of Intelligent Systems, 1993, 8(4): 539-567.

[67] DING S F, LI H, SU C Y, et al. Evolutionary artificial neural networks: a review[J]. Artificial Intelligence Review, 2013, 39(3): 251-260.

[68] YAO X, LIU Y. A new evolutionary system for evolving artificial neural networks[J]. IEEE Transactions on Neural Networks, 1997, 8(3): 694-713.

[69] SALIMANS T, HO J, CHEN X, et al. Evolution strategies as a scalable alternative to reinforcement learning[J]. arXiv preprint, 2017, arXiv:1703.03864.

[70] SUCH F P, MADHAVAN V, CONTI E, et al. Deep neuroevolution: genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning[J]. arXiv preprint, 2017, arXiv:1712.06567.

[71] 栾丽华, 吉根林. 决策树分类技术研究[J]. 计算机工程, 2004, 30(9): 94-96, 105.
LUAN L H, JI G L. The study on decision tree classification techniques[J]. Computer Engineering, 2004, 30(9): 94-96, 105.

[72] 鲁大剑. 面向作战推演的博弈与决策模型及应用研究[D]. 南京: 南京理工大学, 2013.
LU D J. Research on game and decision model for operational deduction and its application[D]. Nanjing: Nanjing University of technology, 2013

[73] 尹星, 孙鹏, 韩冰. 基于决策树的作战实体行为规则建模[J]. 指挥控制与仿真, 2020, 42(1): 15-19.
YIN X, SUN P, HAN B. Modeling of behavior rules of combat entities based on decision tree[J]. Command Control & Simulation, 2020, 42(1): 15-19.

[74] ZHOU Z H, FENG J. Deep forest[J]. National Science Review, 2019, 6(1): 74-86.

[75] 董浩洋, 张永亮, 齐宁, 等. 基于综合势能的作战行动序列生成方法研究[J]. 军事运筹与系统工程, 2020, 34(3): 11-18.
DONG H Y, ZHANG Y L, QI N, et al. Research on the method of generating operational sequence based on comprehensive potential energy[J]. Military Operations Research and Systems Engineering, 2020, 34(3): 11-18.

[76] BREIMAN L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.

[77] DE MESENTIER SILVA F, TOGELIUS J, LANTZ F, et al. Generating novice heuristics for post-flop poker[C]//Proceedings of 2018 IEEE Conference on Computational Intelligence and Games. Piscataway: IEEE Press, 2018: 1-8.

[78] 周献中, 郭庆军, 鞠恒荣. 基于人件服务的 C^4ISR 服务视点扩展[J]. 指挥信息系统与技术, 2016, 7(5): 1-9.
ZHOU X Z, GUO Q J, JU H R. Extended C^4ISR service viewpoint based on humanware service[J]. Command Information System and Technology, 2016, 7(5): 1-9.

[79] 朱咸军, 周献中, 王友发, 等. 面向新型决策系统的人件模型研究[J]. 中国科技论坛, 2016(6): 121-127.
ZHU X J, ZHOU X Z, WANG Y F, et al. Research on humanware model of novel decision system-oriented[J]. Forum on Science and Technology in China, 2016(6): 121-127.

[80] LUCAS Simon, 沈甜雨, 王晓, 等. 基于统计前向规划算法的游戏通用人工智能[J]. 智能科学与技术学报, 2019, 1(3): 219-227.
SIMON L, SHEN T Y, WANG X, et al. General game AI with statistical forward planning algorithms[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(3): 219-227.

[81] SHAO K, ZHU Y H, ZHAO D B. StarCraft micromanagement with reinforcement learning and curriculum transfer learning[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2019, 3(1): 73-84.

[82] SILVER D, HUBERT T, SCHRITTWIESER J, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play[J]. Science, 2018, 362(6419): 1140-1144.

[83] TANG Z T, ZHU Y H, ZHAO D B, et al. Enhanced rolling horizon evolution algorithm with opponent model learning[J]. IEEE Transactions on Games, 2020: 1.

[84] 杨旭, 王锐, 张涛. 面向无人机集群路径规划的智能优化算法综述[J]. 控制理论与应用, 2020, 37(11): 2291-2302.
YANG X, WANG R, ZHANG T. Review of unmanned aerial vehicle swarm path planning based on intelligent optimization[J]. Control Theory & Applications, 2020, 37(11): 2291-2302.

[85] 张菁, 何友, 彭应宁, 等. 基于神经网络和人工势场的协同博弈路径规划[J]. 航空学报, 2019, 40(3): 322493.
ZHANG J, HE Y, PENG Y N, et al. Neural network and artificial potential field based cooperative and adversarial path planning[J]. Acta Aeronautica et Astronautica Sinica, 2019, 40(3): 322493.

[86] LEE D, TANG H R, ZHANG J O, et al. Modular architecture for StarCraft II with deep reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. [S.l.:s.n.], 2018.

[87] MEENAKSHI N. An efficient agent created in StarcCraft 2 using pysc2[J]. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021, 12(10): 336-342.

[作者简介]

孙宇祥（1990– ），男，南京大学工程管理学院博士生，主要研究方向为智能博弈与作战推演。

彭益辉（1995– ），男，南京大学工程管理学院硕士生，主要研究方向为多智能体深度强化学习技术。

李斌（1998– ），男，南京大学工程管理学院硕士生，主要研究方向为分层强化学习及智能博弈。

周佳炜（1997–），男，南京大学工程管理学院硕士生，主要研究方向为深度强化学习算法设计。

张鑫磊（1996– ），男，南京大学工程管理学院硕士生，主要研究方向为智能体多通道人机交互及智能博弈。

周献中（1962– ）男，博士，南京大学工程管理学院教授，主要研究方向为混合智能系统协作与任务规划、指挥与控制系统理论与技术等。