游戏 AI：对智能作战推演的启示

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。不论在游戏 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 如何发展应用到实际的智能作战推演依然面临着巨大的困难。本文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所需的主要属性需求，并结合当前最新的深度强化学习发展概况进行综述。从智能博弈领域主流研究技术、相关智能决策技术、作战推演技术难点三个维度综合分析游戏 AI 发展为智能作战推演的可行性，并最后给出未来智能作战推演发展建议。本文综述了智能博弈领域的整体进展、已有工作以及潜在研究方向，可以为本领域研究人员提供一个较为清晰的发展现状及有价值的研究思路启发。

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

Game AI: Enlighten to intelligent combat deduction

SUN Yuxiang[1], PENG Yihui[1], LI Bin[1], ZHOU Jiawei[1], ZHANG Xinlei[1], ZHOU Xianzhong[1,2]

1.School of Engineering Management, Nanjing University, Nanjing 210093, Jiangsu China;

2.Research Center for New Technology in Intelligent Equipment Nanjing University, Nanjing Jiangsu 210007, China

Abstract: The field of intelligent game has gradually become hotspot of AI research. Both in the field of game AI and intelligent wargame, a series of research breakthroughs have been made in recent years. However, how to develop the game AI and apply it to the actual intelligent combat deduction is still facing great difficulties. This paper comprehensively analyzes the overall research progress in the field of intelligent game at domestic and overseas, analyzes the main technical requirements of intelligent combat deduction in detail, and summarizes the latest development of deep reinforcement learning technology. This paper analyzes the feasibility of developing game AI into intelligent combat deduction from three dimensions of mainstream research technology, relevant intelligent decision-making technology and technical difficulties of combat deduction in the field of intelligent game, and finally gives some suggestions for the development of intelligent operation deduction in the future. This paper summarizes the overall progress, existing work and

potential research direction of intelligent game for the first time in China. It is believed that this paper can provide a clear development status and valuable research ideas for researchers in this field.

Key words: Intelligent game, Game AI, Intelligent combat deduction, Intelligent wargame, Deep reinforcement learning

1 引言

以 2016 年 AlphaGo 的研发成功为起点，智能博弈领域的研究在近五年获得突飞猛进的进展。2016 年之前，兵棋推演的研究还主要集中在基于事件驱动、规则驱动的较为固定的思路研究，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的那么遥远。随着机器学习技术的发展，在一个新的领域——游戏 AI 打开了一个全新的大门，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。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[1]。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[2][3]。然而，随着近五年国内外在各种棋类、策略类游戏领域取得新突破，使智能作战推演的发展迎来的新的机遇[4]。

国内游戏 AI 领域近三年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[5]。网易伏羲实验室在很多游戏环境，如潮人篮球，逆水寒，倩女幽魂都进行了强化学习游戏 AI 的尝试[6]。超参数打造了游戏 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 与游戏场景进行结合，提供人工智能解决方案。启元 AI“星际指挥官”也取得了战胜职业选手的案例[7]。字节跳动也在近期收购了沐瞳科技和深极智能，准备在游戏 AI 领域发力。除了在游戏 AI 领域，国内在智能兵棋推演领域也迅速发展。国防大学兵棋团队研制了战略、战役级兵棋系统，并分析了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[8]。中国科学院自动化所在 2017 年首次推出“CASIA-先知 1.0”兵棋推演人机对抗 AI[9]，并近期上线“庙算·先胜”即时策略人机对抗平台[10]。此外，由中国指挥与控制学会和华戍防务共同推出的专业级兵棋《智戎·未来指挥官》，作为《“墨子”联合作战推演系统》的民用版本，在第三届、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完成了 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、陆军工程大学、中国电科 52 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[11][12][13][14]。在 2020 年，国内举办了四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。国内从游戏 AI 到智能兵棋的发展也在近五年逐渐取得了国内学者的关注，胡晓峰教授撰文提出了从游戏博弈到作战指挥的决策差异，分析了现有主流的人工智能技术在应用到战争对抗过程中的局限性[4]。南京理工大学张振、李琛利用 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[15][16]。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[17][18]。海军研究院和中科院自动化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[19]。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[20]。

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[21]。2015 年 DeepMind 团队在 Nature 上发表了深度 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[22]。2017 年，DeepMind 团队根据深度学习和策略搜索的方法推出了 AlphaGo[23]，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[24]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 Agent 深度强化学习推出的 AlphaStar[25]在 StarCraftII 游戏中达到了人类大师级的水平，并且在 StarCraftII 的官方排名中超越了 99.8%的人类玩家。Dota2 AI “OpenAI Five”在电竞游戏中击败世界冠军[26]，以及 Pluribus 在六人无限制德州扑克中击败人类职业选手[27]。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 Atari 游戏[28]。和军事推演直接相关的 CMANO 和 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[29]。美国兰德公司也对兵棋推演的应用进行相关研究报告，利用兵棋推演假设分析了俄罗斯和北约之间的对抗结果，并利用智能兵棋推演去发现新的战术[30]。兰德研究员也提出利用兵棋来作为美国军事人员学习战术战法的工具[31]。但就目前而言，将机器学习技术应用到作战推演 AI 里还有很多问题需要解决，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，对各类作战智能指挥决策带来翻天覆地的变化。

2 智能作战推演主要属性需求

2.1 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、作战实体所处状态等要

素的表现，状态空间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 中状态空间是在每一帧每个单位可能有的不同状态，如生命值，级别，金币[32]。对于墨子未来指挥官中，主要是每个作战单元实体的状态信息，并且需要把想定中敌我双方所有的作战单元信息汇聚形成。这里状态空间要和可观察空间区分，可观察空间主要为每个 Agent 可以观察到的状态信息，是整个状态空间的一部分。对于作战推演中的状态空间将更加复杂，具有更多的作战单位，更多的单位状态。针对敌我双方的不同作战单位、不同单位的属性、不同的环境属性等定义作战推演的状态空间属性。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。针对陆战环境应包括周围道路信息、城镇居民地、夺控点等。

2.2 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 这类游戏动作主要是玩家控制一个英雄进行一系列的操作，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 玩家有几十个选项来做动作选择，包括有 24 个方向的移动按钮，和一些相应的释放位置/方向的技能按钮[33]。所以每局 MOBA 游戏的动作空间可以达到 10^{60000+}。假设游戏时长 45 分钟、每秒钟 30 帧，共计 80,000 帧，AI 每 4 帧进行一次操作，共计 20,000 次操作，这是游戏长度。任一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（比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等于 1,000，也即动作空间。因此，操作序列空间约等于 $1000^{20000} = 10^{60000}$。而对于星际争霸这类实时策略对抗游戏，因为要控制大量的作战单元和建筑单元，动作空间可以达到 10^{80000+}。而对于 CMANO 和墨子未来指挥官这类更加贴近军事作战推演的游戏，需要对每个作战单元进行大量精细的控制，针对作战推演每个作战单元都应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 $10^{100000+}$。可以看出对于作战推演，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，根据战略决策制定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是离散还是连续，对于 Atari 和围棋这类游戏动作都是离散动作空间[34][23]，而对于星际争霸、CMANO、墨子未来指挥官这类游戏主要还是连续动作[35]。

对于离散动作可以考虑基于值函数的强化学习进行训练，而对于连续的动作空间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，方法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，转化为离散动作。

2.3 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推演平台中体现的较为明显。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，比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。这一决策空间和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，这一阶段的具体微观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，一局想定推演双方博弈决策空间数量预估为 10^{80+}个，而对于星际争霸、Dota2 和王者荣耀这类 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。

2.4 胜利条件设置

不同游戏博弈对抗的胜利条件主要是一局对抗结束的标志，达到目标获得游戏对抗的胜利。而在不同游戏中的胜利条件类型也各不同，在围棋、国际象棋这些棋类博弈对抗过程中有着清晰明确的获胜条件[28]。而对于 Atari 这类游戏[34]，只需要获得足够的分数即可以获得胜利。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，在奖赏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，甚至实际作战推演则有着更加复杂、更加多目标的获胜条件。比如，有些情况可能需要考虑实现我方损失最低的情况下实现作战目标，而有些情况则需要不计代价的快速实现作战目标，这些复杂多元的胜利条件设置将使得强化学习的回报值设置不能是简单的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实现在复杂多变的作战场景中满

足不同阶段、不同目标的作战要求。

2.5 回报值设置

博弈对抗过程中最核心的环节是设置奖赏回报值，合理有效的奖赏值设定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，比如取得最终胜利设置固定的回报值。但是一局游戏中的时间有时较长，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点，即奖励稀疏问题。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的获得回报值或损失回报值的点。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、占领夺控点即获得回报值。如果被打击、失去夺控点等损失回报值，甚至为了加快收敛防止算子长期不能达到有效地点，会在每个 step 都损失微小的回报值。对于王者荣耀的觉悟 AI 也同样设置了详细的奖赏表[33]，从资源、KDA、打击、推进、输赢五个维度制定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，回报函数可能还需要更加细节的制定。因为作战情况将更加复杂多样，需要利用逆强化学习，通过以往的作战数据反向构建奖赏函数。

2.6 战争迷雾

战争迷雾主要是在博弈对抗过程中存在信息的不完全情况，对于未探索的区域并不了解实际的态势信息。对于围棋、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、Dota2、王者荣耀以及 CMANO 等策略 RTS 游戏中设计了这一机制。对于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的博弈问题解决这个问题，已有学者利用不完全信息博弈解决了德州扑克中不完全信息问题[27]，但是在实际作战推演中这一问题还需要进一步探讨研究。

2.7 观察信息

这里智能博弈中的观察信息需要和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。游戏中观察信息以星际争霸、围棋、王者荣耀和 Dota2 为例。星际争霸中观察信息主要有两个层面意思，一个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 也是按照这种限制对星际争霸中的注意力区域进行限制，从而更好的防止人工智能 AI 产生作弊行为。而这也是部分星际争霸 AI 被人诟病所在，即没有限制机器的关注区域。第二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，而对于不能获取的态势信息则只能评估预测，而这一部分则涉及到对手建模部分，这一部分则主要利用 POMDPs[36]（部分可观测马尔可夫决策过程），这一技术明显难于完全信息博弈，在下一小节详细介绍。而另一个极端则是围棋游戏中的信息完全可获取，属于完全信息博弈，态势信息即等于观察信息，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，对于围棋游戏中的观察信息则可以利用 MCTS 方法对所获取的信息进行详细分析，计算出所有可能的结果，进而得出最佳的方案策略。对于 Dota2 中的观察信息是对所控制的某个英雄所获取的态势信息进行处理，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。这一部分腾讯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后，为战略方案提供支持，如英雄是去野区发育还是去中路对抗。对于主屏幕信息获取后则主要对当前态势信息训练后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，亦或者推塔进攻。

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、舰艇雷达等各方面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。所以，对于作战推演观察信息完全可以借鉴 POMDPs 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针对性的条件，如作战飞机偷袭对方时使用铝箔条干扰、降低飞行高度等手段防止敌方发现，这些特殊战术规则需要提前完整制定才能更加真实的获取作战推演信息，这是游戏推演中差距甚远的。

2.8 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，并且也直接影响推演对抗的价值[37]。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大。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。而在中科院自动化所开发的庙算·智胜上积分排名前三名均是 AI 选手，胜率均在 80%以上[10]。但是，现有对手建模主要还是聚焦在 1vs1 的对手建模，很少有研究在三方博弈或者更多的博弈方的情况下，如何进行多方博弈，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”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。

2.9 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，想定可以说是永久固定不变的，而且也完全抽象了环境的影响，所以 AlphaGo 这类智能 AI 完全没有考虑环境的因素。在觉悟 AI、Dota2 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。不同的英雄搭配实际也是不同的环境，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就要开始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，这些因素对于作战推演的过程会产生非常重要的影响。CMANO、墨子未来指挥官、Wargame：Red Dragon 中都需要考虑地形因素。比如 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，这在先知兵圣早期的比赛中就比较突出的显示了这一问题。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，更换作战想定即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和先验知识+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。

2.10 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[5、7、23、25、26、28、29、38-41]进行分析，对比不难发现游戏 AI 过渡到智能兵棋，以至于到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，具体如表 1 所示。

表 1 各博弈环境关键属性对比

游戏/兵棋	状态空	动作空间	决 策	胜 利	回 报	战 争	观 察	对 手	想 定

	间		数量	条件	值设置	迷雾	信息	建模	设计
Go	中等	中等	中等	数子法/数目法	简单	无	简单	中等	固定
StarcraftⅡ	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
Wargame：Red Dragon	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

3 作战推演的智能决策核心技术思路

3.1 强化学习技术框架

强化学习的核心思想是不断在环境中探索试错，并得到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。马尔可夫过程（MDP）是强化学习的基础模型，环境通过状态与动作建模，描述智能体与环境的交互过程。一般地，MDP 可表示为四元组$< S, A, R, T >$[42]：

- S 为有限状态空间（State Space），包含 Agent 在环境中所有的状态；
- A 为有限动作空间(Action Space)，包含了 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s状态下，执行a动作，到达下一状态s'，Agent 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)，$P_{ss'}^{a} = P[S_{t+1} = s' \mid S_t = s, A_t = a]$表示在状态$s$上执行动作$a$，并转移到状态$s'$的概率。

在 MDP 中，Agent 与环境交互如图 1 所示

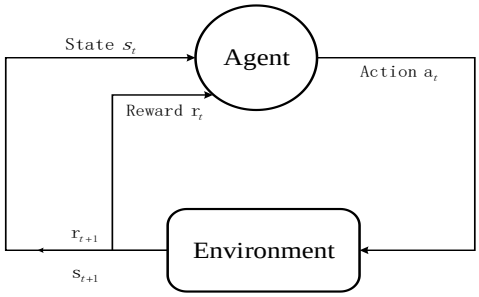

图 1 Agent 与环境交互

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，环境给以 Agent 相应的奖赏信号反馈r_{t+1}，并以一定概率转移到新的环境状态s_{t+1}，等待 Agent 做出下一步新的决策。在与环境的交互过程中，Agent 有两处不确定性产生，一处是在状态 S 处选择什么样的动作，用策略$\pi(a|s)$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积奖赏，即

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

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 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$Q^*(s，a)$公式如下；

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

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

3.2 强化学习主流算法

3.2.1 值函数强化学习

强化学习最早利用 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，它有一张 Q 表来保存所有的 Q 值，当动作空间巨大的时候，该算法就难以适应。所以，Q-Learning 算法只能在比较简单的环境中建模使用，如简单的迷宫问题中，让 Agent 通过 Q-Learning 算法自动寻找出口。

DeepMind 在 2015 年第一次利用 DQN 算法在 Atari 环境中实现了高水平的智能 AI，该 AI 综合评定达到人类专业玩家的水平[34]。这也使得 DQN 算法成为了强化学习的经典算法。

该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，进而获得精准的预测 Q 值，并通过最大的 Q 值来进行动作选择。DQN 算法有效的解决了 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 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，所以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，因此使得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 主要应用于离散的动作空间，且 DQN 的训练不一定能保证 Q 值网络收敛，这就会导致在情况过于复杂的情况下，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 算法[43]、Prioritized Replay DQN 算法[44]、Dueling DQN 算法[45]等。这些算法主要是改进 Q 网络防止过拟合、改进经验回放中的采样机制、优化目标 Q 值计算等方面提升传统 DQN 网络的性能。总的来说 DQN 系列强化学习算法都属于值估计（Value Based）的强化学习算法类型。基于值估计的强化学习算法主要存在以下三点不足。

- 对连续动作的处理能力不足。DQN 之类的方法一般都是只处理离散动作，无法处理连续动作。
- 对受限状态下的问题处理能力不足。在使用特征来描述状态空间中的某一个状态时，有可能因为个体观测的限制或者建模的局限，导致真实环境下本来不同的两个状态却在建模后拥有相同的特征描述，进而很有可能导致 Value Based 方法无法得到最优解，此时使用策略函数（Policy Based）强化学习方法也很有效。
- 无法解决随机策略问题。Value Based 强化学习方法对应的最优策略通常是确定性策略，因为是从众多行为价值中选择一个最大价值的行为，而有些问题的最优策略却是随机策略，这种情况下同样是无法通过基于价值的学习来求解的。这时也可以考虑使用 Policy Based 强化学习方法。

由于上面这些原因，Value Based 强化学习方法不能适用所有的场景，因此需要新的解决上述类别问题的方法，比如 Policy Based 强化学习方法。

3.2.2 策略函数强化学习

在 Value Based 强化学习方法里，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，并接受状态 s 与动作 a 作为输入，计算后得到近似的动作价值，即：

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

在 Policy Based 强化学习方法下，主要采用类似的思路，只不过这时候主要是对策略进

行近似表示。此时，策略可以被描述成为一个包含参数 θ 的函数，θ 主要为神经网络中的权重，即：

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

在 Policy Based 强化学习方法中，最经典的就是 Sutton 在 2000 年提出的 AC 框架强化学习算法。Actor-Critic 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 使用我们之前讲到了的价值函数，负责评估 Actor 的表现，并指导 Actor 下一阶段的动作。总的来说，Critic 通过 Q 网络计算状态的最优价值 V_t，而 Actor 利用 V_t 这个最优价值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 来帮 Actor 计算状态的最优价值 V_t。

在 2016 年 DeepMind 在 ICML 提出了 A3C 算法[46]。之前的 DQN 算法，为了方便收敛使用了经验回放的技巧；Actor-Critic 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。A3C 利用多线程的方法，同时在多个线程里面分别和环境进行交互学习，每个线程都把学习的成果汇总起来，整理保存在一个公共的地方。并且，定期从公共的地方把所有 Actor 的学习成果拿回来，指导自己和环境后面的学习交互。具体的改进是提出了异步训练框架即 Global Network。该部分作为共享的公共部分，主要是一个公共的神经网络模型，这个神经网络包括 Actor 网络和 Critic 网络两部分的功能。下面有 n 个 worker 线程，每个线程里有和公共的神经网络一样的网络结构，每个线程会独立的和环境进行交互得到经验数据，这些线程之间互不干扰，独立运行，n 个线程会独立的使用累积的梯度分别更新公共部分的神经网络模型参数。A3C 解决了 Actor-Critic 难以收敛的问题，同时更重要的是，提供了一种通用的异步并发强化学习框架，也就是说，这个并发框架不光可以用于 A3C，还可以用于其他的强化学习算法。这是 A3C 最大的贡献。同时 A3C 算法作为可以处理连续动作空间的算法，进行了大量实验，并且提出了各种进一步提升效率的途径，如更好的利用 GPU 训练 A3C 算法等。

确定性策略是和随机策略相对而言的，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，这样动作的空间维度极大。如果我们使用随机策略，即像 DQN 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才可行的。于是 DeepMind 就想出使用确定性策略来简化这个问题[47]。作为 DDPG，Critic 目标网络和 DDQN 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’，这部分工作由于用来估计目标 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，加一些其他的优化，比较好的解决了 Actor-Critic 难收敛的问题。因此在实际产品中尤其是自动化相关的产品中用的比较多，是一个比较成熟的 Actor-Critic 算法。后期 OpenAI 在 2017 年 NIPS 又提出了改进的 MADDPG 算法[48]，把强化学习算法进一步推广应用多 Agent 环境，下面小节会进一步介绍。在 AC 框架下，比较经典的算法还有 PPO 算法[49]、SAC 算法[50]、TD3[51]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。

3.3 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。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，对于主视野和小地图都采取了不同种类的要素提取到一层，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[5][33]。最终将多尺度特征的信息融合形成全局态势特征信息，这一工作在 AlphaStar 中也是这样利用的。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，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技术，然后将提取信息与语义模型相结合，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，这样就可以让强化学习、深度学习、自然语言处理进行融合，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。

3.4 分层强化学习

在智能博弈对抗的建模过程中，面临着两个难题，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，Open AI 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 low-level 动作组成 high-level 动作，这样搜索空间就会被降低[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，和要解决的对抗任务的关系不是十分强。接下来学习一个高层的控制策略能够使智能体根据状态调用这些技能，能够很好地解决探索问题，可以在一系列稀疏奖励的任务中表现出色[53][54]。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 1 级和 2 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，它告诉我们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，是哪个英雄，还是小兵，还是防御塔等；如果第一个层级是技能 1（或 2，或 3），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着不同的动作，比如坦克可以选择直瞄射击、间瞄射击，移动方向等，对于实际作战推演不同装备同样有众多复杂的动作，通过这样的特征和标签设计，人工智能建模任务可以作为一个层次化的多类分类问题来解决。具体来说，一个深层次的神经网络模型被训练来预测在给定的情境下要采取什么行动。对于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[33]。南京大学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[55]。分层强化学习比较通用的框架是分为两层，顶层策略为 meta-controller 负责生成总体宏观目标，底层策略称为 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、战术三个层次，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[56]。

3.5 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。在这一方面 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 都有一个相同的的训练神经网络，但是没有一个全局控制网络[26]。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 都建立属于自己的神经网络进行训练。最后一种方法是最理想的状态，但是训练过程复杂，也难以适用于大规模的推演过程[25]。对于实际作战推演，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，还需要重新训练一遍网络，

这样考虑的话让每一个 Agent 有自己的神经网络应该是更加好的选择。

3.6 LSTM 技术结合深度强化学习

在觉悟 AI 的设计中，利用了深度学习不断提取游戏界面的态势信息。但是，利用深度学习技术虽然可以把一个对抗界面的所有特征提取出来了，但是提取的是一个静态的某一帧的界面信息，并没有把时间步与时间步之间的信息关联起来。所谓一个时间步一般指一帧，当然可以指多帧，那么如何把前面历时的帧信息和现在的信息关联起来？这里就引入了 LSTM（Long Short-Term Memory）长短期记忆网络。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点应该在主画面和小地图的哪个方面，进而综合输出合理的动作，并且也通过 LSTM 关联历史数据来训练强化学习的神经网络模型[57]。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。

3.7 多属性决策结合强化学习

强化学习的回报值设置往往根据专家经验手工设置，但是这种手工设置回报值往往难以确定具体正确的估计值，且训练长时间才能评估回报值设置的好坏。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进行排序，计算对应算子对我方的威胁度，通过计算的威胁度和强化学习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，这样有利于科学提高强化学习训练的智能性，并有利于加快收敛。

4 其他可用智能决策技术

在智能博弈领域，国际上 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 AI、CASIA-先知也都取得了突破性的进展。这些工作都主要以深度强化学习技术为主，但均都搭配使用了其他相关的人工智能技术。总的来说，单纯的利用深度强化学习技术并不能很有效的实现智能 AI，有必要在训练过程中结合其他的技术完善提高 AI 性能。同时，如果想要实现特别突出的 AI 智能，训练的过程中也需要大量的成本。AlphaStar 的训练过程持续了十个月，使用了 51000CPUs，并且同时有 30 个博士生参与工作，成本达到百万美元，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[40]。

4.1 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，并逐步淘汰掉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，但是在当时计算机算力不足，导致这个方向并没有引起过多关注[58][59][60]。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替代强化学习技术，在 MuJoCo 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[61]。Uber AI 在 2017 年尝试利用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，取得的算法性能一定程度优于 A3C、DQN 算法[62]。除了在优化网络参数方面进行结合，将进化策略和多智能体强化学习方面进行结合，也是一个有意义的方向。DeepMind 在 AlphaStar 中就是利用了联盟赛制，从而在训练出的 Agent 中不断优化筛选出更加优秀的 Agent，进行不断演化出最终超过职业选手水平的游戏 AI[25]。

总的来说，单纯利用遗传算法相比较于强化学习算法有着明显的缺陷，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。在实际推演中就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，遗传算法的优势就是适合在大规模的空间探索中，寻找全局最优解。而强化学习算法随着梯度下降进行优化，很容易寻找到的是局部最优解，而不是全局最优解。所以，如果找到合适的结合角度，相信进化算法和强化学习两者结合可以产生一定有价值的工作。

4.2 决策树

决策树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[63]。其本身是一种树形结构，其中每个内部节点表示一个属性上的判断，每个分支代表一个判断结果的输出，最后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[64]，其易于构建作战实体的行为规则建模，有利于分析基于决策树的作战实体行为模型[65]，这对于作战推演的初期快速建立对手模型不失为一种高效办法，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，如南京大学周志华团队提出的深度森林算法[66]，就是在决策树的基

础上拓展而来。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。

4.3 基于规则

基于规则的 AI 主要是结合博弈对抗环境的领域知识，构建出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，通过程序化形成具有一定智能性的推演 AI。基于规则的 AI 在国内研究开展较早，但是直到近年来的实际应用中，还是比较普遍以规则 AI 为基础进行改良和设计[3]。国内在 2020 年的智能兵棋比赛中，大部分团队及 baseline 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，可以快速构建具体一定智能性的博弈对抗 AI 环境。可以作为对手模型构建，强化学习 AI 与基于规则的 AI 对抗，来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，来让 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[17]。

4.4 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[67]。该方法实际是综合利用离线和在线统计数据进行分析，综合分析出智能兵棋推演 AI。该算法可以尝试跟强化学习结合，弥补强化学习开始阶段训练收敛速度过慢，并可以在强化学习算法执行过程中，考虑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。

4.5 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。Leo Breiman 和 Adele Cutler 推论出随机森林的算法[68]。 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，子数据集的数据量是和原始数据集相同的。并且待选特征可以进行随机选取，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，之后再在随机选取的特征中选取最优的特征。其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，且不容易过拟合。对于高维数据，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随机森林相关研究其实较少，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问题。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[69]。

4.6 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到人机交互系统中，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[70][71]。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，主要是利用高水平的玩家数据进行监督学习，方便快速高效的训练出高水平的深度强化学习 AI。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。强化学习算法可以直接从已有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，结果在没有人类数据的强化学习中，复杂环境中很难有效果。在仅仅有监督学习技术的支持下，可以达到不错的效果。在充分利用人类数据后，AlphaStar 的性能可以再次提高 60%。在对于高水平玩家经验数据的使用方面，AlphaStar 主要在以下几个方面给出参考:

- 监督学习预训练模型；
- 用高水平玩家数据约束探索，缩小行为空间；
- 利用人类数据构建奖励函数，引导策略模仿高水平玩家行为。对于智能作战推演，也可以考虑充分调动人的行为因素，融合行为决策和强化学习理论，进而建立起高水平的智能作战推演 AI。

4.7 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。常见的经典模型为蒙特卡洛树搜索（Monte Carlo Tree Search，MCTS）模型方法，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[23]这样的游戏，分支因子要大几个数量级，而有用的启发式又很难形成，这类问题就需要用 MCTS 算法来解决[72]。尽管 MCTS 算法在大范围的博

弈中提供了更强的决策能力，但其应用在作战推演领域仍存在着很多挑战和瓶颈。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 算法是否仍然为指导作战推演的最佳方法是一个有待研究的问题。

4.8 小地图设置技术

在多个智能博弈对抗游戏中，都普遍存在着一个小地图，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，指挥员可能关注在某个局部作战场景中，同时也应该考虑全局的作战信息的获取。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。

5 作战推演技术难点及技术解决方案

5.1 小地图设置技术

研究人员在对强化学习的训练过程中总是会遇到强化学习训练过程时间长，难以收敛的问题，这种情况通俗称为冷启动问题。针对这个问题，现有研究人员提出了多种解决方案，最为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 experience memory 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专家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。

5.2 过拟合问题

在智能博弈对抗过程中经常会产生训练一定阶段后，就陷入局部最优结果。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，进而称为产生训练的过拟合现象。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，对于一定比例的动作选择概率下的随机探索，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，而不是固定唯一不变的回报。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，从而避免了因为固定对手而导致的过拟合现象。

5.3 想定适应性问题

智能博弈的 AI 建模普遍存在着适应性不高的问题，有部分研究人员开发的 AI 是针对某个固定想定开发，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，所以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。中科院自动化所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[73]。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[74]。觉悟 AI 引入了课程学习（The flow of curriculum self-play learning）方法，将各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[5][33]。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，也只有这样可以更加适应实时性要求极高的作战场景。

5.4 智能蓝方建模

对手建模通常是在两个 Agent 博弈的环境中，为了获得更高的收益，需要对对手（队友）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进行决策。对于智能蓝方建模主要是在具有战争迷雾的情况下，针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，在无法获取到准确的对手信息的情况下，针对对方进行预测评估。其中一种假设对手是完全理性的，针对对手（队友）建模是为了寻找博弈中的纳什均衡策略。为解决这一难点问题，阿尔伯塔大学研究人员提出了（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，建模方式可将对手建模分为隐式建模和显式建模两类。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处理对手信息缺失的问题，通过最大化智能体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[75]。总的来说，对手建模技术是智能博弈对抗是否有效的关键所在，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。

5.5 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[76]。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一个最佳路径。但是，在已有的算法中主要以 A-Star 算法、Dijkstra 算法、D*算法、LPA*算法、D* lite 算法等作为路径规划，在物流运输、无人驾驶、航空航天等领域都取得了显著的成效。同时也有学者提出其他的一些路径规划算法，如基于神经网络和人工势场的协同博弈路径规划方法[77]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。

6 作战推演发展建议

6.1 智能作战推演通用框架

在现有的游戏平台中也有比较成熟的 AI 开发通用框架，如 pysc2[78][79]。但是相比较于成熟的作战推演通用框架还是有较大差距。智能作战推演系统可以设计一个适用于复杂环境中的通用框架，该框架包括作战推演算子、地图、规则、想定。同时最为关键的是设计通用的算法接口，通过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。其中最为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，把智能算法的设计开发和作战推演环境进行解耦，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，可以直观的提供一个通用接口进行展现，作为指挥人员直观理解智能算法给出的建议。

6.2 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去应用。现阶段主要想定应用的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。在墨子未来指挥官中的对抗想定更接近与战役层面的智能决策方案，对于战略层面的智能决策现阶段还较少有研究。其原因就在于所面临的想定越宏观，智能决策所面临的技术挑战越大，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达到成熟的研究阶段。所以，当前更易于考虑战术层面的智能决策。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，以及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

演中考虑进来后勤机制的影响。另外对于战术、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，这一技术思路实现较为简单，所需算力也较低，可以考虑为初期实现的路径之一。

6.3 人机融合的智能作战推演模式建立

对于智能作战推演的未来趋势主要分为人在环和人不在环两种类型。对于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。另一种人在环的模式又分为两种，一种是实现人机对抗，国内已有这方面比赛，通过开发训练好的智能算法 Agent，通过高水平指挥人员与之进行对抗，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，对于人机融合模式的框架研究也需要进行探索，如何将人类领域知识引入智能算法中，帮助智能算法更为高效的实现智能作战推演。

6.4 开放性的仿真实验平台建立

随着智能博弈近年的兴起，当前不论是游戏智能 AI 还是智能兵棋平台，国内外高校、研究所、企业都已逐渐开发完成各种类型的平台。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，如果想要尝试智能算法开发，一旦尝试新的平台就需要开发人员重新学习适应新的平台接口和架构，这浪费了大部分研究人员精力。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，但实际上本质都一样，很有必要构建一个通用一体化智能博弈平台框架，防止不断重新开发、学习的过程，提高智能博弈平台的研究效率也是势在必行。

7 结束语

本文尝试详细具体地梳理出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，并和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，期望能为从事游戏 AI、智能兵棋、智能作战推演等智能博弈领域的研究和开发人员提供一定的研究思路。当前智能博弈的研究思路还是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。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较于实际应用还有较大差距，但相信智能博弈这一研究方向一定是未来智能决策研究发展的必由之路，也一定会最终在各个相关领域得以实现。

参考文献：

[1] 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.

[2] 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.

[3] 谭鑫.基于规则的计算机兵棋系统技术研究[D].国防科学技术大学,2010.

[4] 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(04):356-363.

[5] Ye D, Chen G, Zhao P, 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.

[6] Fu H, Tang H, Hao J, et al. Deep multi-agent reinforcement learning with discrete-continuous hybrid action spaces[J]. arXiv preprint arXiv:1903.04959, 2019.

[7] Wang X, Song J, Qi P, et al. SCC: an efficient deep reinforcement learning agent mastering the game of StarCraft II[J]. arXiv preprint arXiv:2012.13169, 2020.

[8] 周超,胡晓峰,郑书奎,夏荣祥.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(01):19-26.

[9] 黄凯奇,兴军亮,张俊格,倪晚成,徐博.人机对抗智能技术[J].中国科学:信息科学,2020,50(04):540-550.

[10] 中国科学院自动化研究所. 即时策略人机对抗平台 [EB/OL]. (2020-11-1) [2020-11-13]. http://wargame.ia.ac.cn Institute of Automation， Chinese Academy of Sciences. Real-time strategy man-machine confrontation platform. [EB/OL]. (2020-11-1) [2020-11-13] .

[11] Sun, Y.; Yuan, B.; Zhang, T.; Tang, B.; Zheng, W.; Zhou, X. Research and Implementation of Intelligent Decision Based on a Priori Knowledge and DQN Algorithms in Wargame Environment. *Electronics* **2020**, *9*, 1668. https://doi.org/10.3390/electronics9101668

[12] 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(02):1-6.

[13] 孙彧,李清伟,徐志雄,陈希亮.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(02):16-20.

[14] 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.

[15] 张振,黄炎焱,张永亮,陈天德.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(01):77-83.

[16] 李琛,黄炎焱,张永亮,陈天德.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(03):755-762.

[17] 程恺,陈刚,余晓晗,刘满,邵天浩.知识牵引与数据驱动的兵棋 AI 设计及关键技术[J/OL].系统工程与电子技术: 1-10[2021-06-04]. http://kns.cnki.net/kcms/detail/11.2422.TN.20210416.1106.002.html.

[18] 张可,郝文宁,余晓晗,靳大尉,邵天浩.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.

[19] 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(09):116-121.

[20] 施伟,冯旸赫,程光权,黄红蓝,黄金才,刘忠,贺威.基于深度强化学习的多机协同空战方法研究[J/OL].自动化学报:1-16[2021-06-04].https://doi.org/10.16383/j.aas.c201059.

[21] Wang HN, Liu N, Zhang YY, et al. Deep reinforcement learning: a survey[J]. Frontiers of Information Technology & Electronic Engineering, 2020, 21(12): 1726-1744.

[22] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[23] Silver D, Huang A, Maddison CJ, et al. Mastering the game of go with deep neural networks and tree search. Nature, 2016, 529(7587): 484-489.

[24] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[25] Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster level in StarCraftII using multi-agent reinforcement learning[J]. Nature, 2019, 575(7782): 350-354.

[26] Berner C, Brockman G, Chan B, et al. Dota 2 with large scale deep reinforcement learning[J]. arXiv preprint arXiv:1912.06680, 2019.

[27] Brown N, Sandholm T. Superhuman AI for multiplayer poker[J]. Science, 2019, 365(6456): 885-890.

[28] 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.

[29] Price M. What Impact do VR Controllers Have on the Traditional Strategy Game Genre[D]. University of Huddersfield, 2019.

[30] Shlapak, David A. and Michael Johnson, Reinforcing Deterrence on NATO's Eastern Flank: Wargaming the Defense of the Baltics. Santa Monica, CA: RAND Corporation, 2016. https://www.rand.org/pubs/research_reports/RR1253.html.

[31] Tarraf, Danielle C., J. Michael Gilmore, D. Sean Barnett, Scott Boston, David R. Frelinger, Daniel Gonzales, Alexander C. Hou, and Peter Whitehead, An Experiment in Tactical Wargaming with Platforms Enabled by Artificial Intelligence. Santa Monica, CA: RAND Corporation, 2020. https://www.rand.org/pubs/research_reports/RRA423-1.html.

[32] Ye D, Liu Z, Sun M, et al. Mastering complex control in moba games with deep reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(04): 6672-6679.

[33] Ye D, Chen G, Zhang W, et al. Towards playing full moba games with deep reinforcement learning[J]. arXiv preprint arXiv:2011.12692, 2020.

[34] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning[J]. arXiv preprint arXiv:1312.5602, 2013.

[35] Risi S, Preuss M. Behind DeepMind’s AlphaStar AI that Reached Grandmaster Level in StarCraft II[J]. KI-Künstliche Intelligenz, 2020, 34(1): 85-86.

[36] Silver D, Veness J. Monte-Carlo planning in large POMDPs[C]. Neural Information Processing Systems, 2010.

[37] Goodman J, Lucas S. Does it matter how well I know what you’re thinking? Opponent Modelling in an RTS game[C]//2020 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2020: 1-8.

[38] Michael Johanson. Measuring the Size of Large No-Limit Poker Games[J/OL].https://poker.cs.ualberta.ca/publications/2013-techreport-nl-size.pdf, 2013.

[39] OpenAI.OpenAI Five[EB/OL]. https://openai.com/blog/openai-five/, 2018

[40] Vinyals, Oriol and Babuschkin, et al. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II [EB/OL]. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/, 2019. StarCraft II Demonstration. https://www.youtube.com/watch?v=cUTMhmVh1qs

&t=1636s

[41] Wikipedia: Game complexity https://en.wikipedia.org/wiki/G

[42] Sutton R S, Barto A G. Reinforcement learning: An introduction[M]. MIT press, 2018.

[43] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2016, 30(1).

[44] Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay[J]. arXiv preprint arXiv:1511.05952, 2015.

[45] Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning[C]//International conference on machine learning. PMLR, 2016: 1995-2003.

[46] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning[C]//International conference on machine learning. PMLR, 2016: 1928-1937.

[47] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[J]. arXiv preprint arXiv:1509.02971, 2015.

[48] Lowe R, Wu Y, Tamar A, et al. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments[C]//NIPS. 2017.

[49] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv:1707.06347, 2017.

[50] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//International Conference on Machine Learning. PMLR, 2018: 1861-1870.

[51] Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods[C]//International Conference on Machine Learning. PMLR, 2018: 1587-1596.

[52] Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods[C]//International Conference on Machine Learning. PMLR, 2018: 1587-1596.

[53] Florensa C, Duan Y, Abbeel P. Stochastic neural networks for hierarchical reinforcement learning[J]. arXiv preprint arXiv:1704.03012, 2017.

[54] Rafati J, Noelle D C. Learning representations in model-free hierarchical reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 10009-10010.

[55] Pang Z J, Liu R Z, Meng Z Y, et al. On reinforcement learning for full-length game of

starcraft[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 4691-4698.

[56] Li S, Wang R, Tang M, et al. Hierarchical reinforcement learning with advantage-based auxiliary rewards[J]. arXiv preprint arXiv:1910.04450, 2019.

[57] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[58] Yao X. A review of evolutionary artificial neural networks[J]. International journal of intelligent systems, 1993, 8(4): 539-567.

[59] Ding S, Li H, Su C, et al. Evolutionary artificial neural networks: a review[J]. Artificial Intelligence Review, 2013, 39(3): 251-260.

[60] Yao X, Liu Y. A new evolutionary system for evolving artificial neural networks[J]. IEEE transactions on neural networks, 1997, 8(3): 694-713.

[61] Salimans T, Ho J, Chen X, et al. Evolution strategies as a scalable alternative to reinforcement learning[J]. arXiv preprint arXiv:1703.03864, 2017.

[62] 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 arXiv:1712.06567, 2017.

[63] 栾丽华,吉根林.决策树分类技术研究[J].计算机工程,2004(09):94-96+105.

[64] 鲁大剑. 面向作战推演的博弈与决策模型及应用研究[D].南京理工大学,2013.

[65] 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(01):15-19.

[66] Zhou Z H, Feng J. Deep forest[J]. arXiv preprint arXiv:1702.08835, 2017.

[67] 董浩洋,张永亮,齐宁,周志宇.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(03):11-18.

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

[69] De Mesentier Silva F, Togelius J, Lantz F, et al. Generating novice heuristics for post-flop poker[C]//2018 IEEE Conference on Computational Intelligence and Games (CIG). IEEE, 2018: 1-8.

[70] 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(05):1-9.

[71] 朱咸军,周献中,王友发,王志鹏.面向新型决策系统的人件模型研究[J].中国科技论坛,2016(06):121-127.

[72] LUCAS Simon,沈甜雨,王晓,张杰.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(03):219-227.

[73] Shao K, Zhu Y, Zhao D. Starcraft micromanagement with reinforcement learning and curriculum transfer learning[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2018, 3(1): 73-84.

[74] 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.

[75] Tang Z, Zhu Y, Zhao D, et al. Enhanced Rolling Horizon Evolution Algorithm with Opponent Model Learning[J]. IEEE Transactions on Games, 2020.

[76] 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.

[77] 张菁,何友,彭应宁,李刚.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(03):228-238.

[78] Lee D, Tang H, Zhang J, et al. Modular architecture for starcraft ii with deep reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. 2018, 14(1).

[79] Meenakshi N. An Efficient Agent Created In Starcraft 2 Using Pysc2[J]. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021, 12(10): 336-342.

[作者简介]

孙宇祥（），男，南京大学工程管理学院博士研究生，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研究生，主要研究方向为基于人工智能技术的智能推演软件研发。

李斌（1998- ） 男，南京大学工程管理学院硕士研究生，主要研究方向为基于人工智能技术的智能推演软件研发。

周佳炜（1997- ） 男，南京大学工程管理学院硕士研究生，主要研究方向为基于人工智能技术的智能推演软件研发。

张鑫磊（1998- ） 男，南京大学工程管理学院硕士研究生，主要研究方向为智能体多通道人机交互。