人工智能与“星际争霸”：多智能体博弈研究新进展

张宏达[1, 2, 3]，李德才[1, 2]，何玉庆[1, 2]

（1. 中国科学院沈阳自动化研究所机器人学国家重点实验室，沈阳 110016；
2. 中国科学院机器人与智能制造创新研究院，沈阳 110016；3. 中国科学院大学，北京 100049）

摘　要：多智能体博弈游戏具有实时对抗、群体协作、非完全信息博弈、庞大的搜索空间、多复杂任务和时间空间推理等特点，是当前人工智能领域极具挑战的难题。同时，该领域研究成果在社会管理、智能交通、经济、军事等领域有广阔的应用前景。以具有代表性的多智能体博弈游戏“星际争霸”为主要研究对象，通过分析研究难度、总结研究方法、介绍研究环境及数据集与竞赛资源，对近年来该领域人工智能研究成果进行了梳理和总结，并对该领域未来可能的发展方向进行预测，为相关研究工作的开展提供可借鉴参考信息。

关键词：多智能体；实时策略；人工智能；对抗博弈；深度强化学习

中图分类号：TP18　　**文献标识码：**A　　**文章编号：**2096-5915（2019）01-0319-12

Artificial Intelligence and StarCraft: New Progress in Multiagent Game Research

ZHANG Hongda[1, 2, 3], LI Decai[1, 2], HE Yuqing[1, 2]

（1. State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China;
2.Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110016, China;
3. University of Chinese Academy of Sciences, Beijing 100049, China）

Abstract: Multiagent games have the characteristics of real-time confrontation, group collaboration, incomplete information game, huge search space and multi-complex tasks, spatiotemporal reasoning, etc. It is a very challenging problems in the field of artificial intelligence. At the same time, the research results in this field have broad application prospects in the fields of social management, intelligent transportation, economy, and military. We take the representative multiagent game StarCraft as the main research object, and analyzes the research results of artificial intelligence in this field by analyzing research difficulty, summarizing research methods, introducing research environment and data sets and competition resources. And summarizing, predicting the future development direction of the field, and providing reference information for the development of related research work.

Keywords: Multiagent; Real-Time Strategy; Artificial Intelligence; Game; Deep Reinforcement Learning

收稿日期：2018-12-19；修回日期：2019-01-10

基金项目：国家自然科学基金（91748208）；国家自然科学基金联合基金（U1608253）；中国科学院联合基金（6141A01061601）

1 引 言

2016年3月，Deepmind科研团队的围棋程序AlphaGo[1]以4∶1的成绩战胜韩国围棋世界冠军李世石，这一研究成果在全球范围内引起巨大轰动，人工智能研究再一次吸引了世界的目光。在攻克围棋这一艰巨任务之后，Deepmind将研究重点转向更加复杂的领域——多智能体博弈游戏，并与美国电子游戏公司暴雪娱乐（Blizzard Entertainment）合作，在星际争霸II的游戏环境基础上开发了可进行更高水平人工智能研究的学习环境。鉴于该领域聚集了当前人工智能研究领域最具挑战的难题，国内外众多科研单位也竞相投入到这一领域当中。多智能体博弈游戏不仅在人工智能研究领域极具研究价值，其在社会管理、智能交通、经济、军事等领域同样具有巨大的潜在应用价值。

对于当前状态或动态变化既无完美信息又无完整信息可用的复杂动态环境，给人工智能研究带来显著挑战[2]。现实社会中很多大型、复杂的动态环境问题如路面交通系统、气象预报、经济预测、智慧城市管理、军事决策等均是实例。然而，对这些实际问题进行建模仿真存在很大困难。与此同时，一系列实时策略游戏提供了与真实环境相似的、非完美和非完整信息、长远规划、复杂问题决策的仿真环境。这些实时策略游戏环境既能模拟现实问题的关键难点，又具有可准确评估、迭代迅速、便于交互和布署、可重复等特点，为解决实际问题提供了绝佳的研究平台。因此，基于实时策略游戏环境的研究工作对人工智能技术的发展和解决复杂的实际问题都有重要意义。在众多的研究平台中，星际争霸以其丰富的环境信息、逼真的环境场景等特点成为常用的理论研究和方法验证平台。

实时策略游戏——星际争霸具有实时对抗、巨大的搜索空间、非完全信息博弈、多异构智能体协作、时空推理、多复杂任务、长远全局规划等特点，同时这些也是人工智能领域极具挑战的难题。自星际争霸第一版游戏于1998年正式发布以来，不少研究者将其作为人工智能研究环境进行了大量的研究。2010年开始，一些星际争霸人工智能游戏程序国际竞赛开始举办，大量人工智能研究和应用成果开始发布。2016年开始，深度学习在星际争霸中的应用展现出强大的信息处理和决策能力，自此之后更多的深度学习和深度强化学习算法被应用到该研究领域。基于星际争霸进行的一系列人工智能研究极大促进了机器学习、深度学习、博弈论、多智能体协作策略等研究领域的发展，对与星际争霸相关的研究成果进行总结，特别是近两年产生的新的研究理论和成果进行梳理，有助于把握该研究领域的研究进展和动向，为与该领域相关的研究提供参考。

综上，本文主要开展了以下几方面的工作。首先介绍星际争霸游戏环境并分析其给人工智能研究所带来的挑战。接着，对现阶段星际争霸相关研究单位研究成果进行介绍，并对该领域的相关研究方法进行了分类。在此基础上，列举了与星际争霸人工智能研究相关的资源，包括研究平台、数据集以及自主游戏程序竞赛。最后，对星际争霸相关领域未来可行的研究方向进行了预测。

2 星际争霸和人工智能

2.1 实时策略游戏——星际争霸

星际争霸是暴雪娱乐公司发布的一款极为经典的多角色实时策略游戏，目前主要有两版。自主游戏程序竞赛基于1998年发行的第一版游戏环境，如图1。2010年发行的第二版游戏以其更为细致逼真的游戏环境和新的竞技模式更受玩家的欢迎，如图2。

星际争霸提供三种类型的角色供玩家选择：人族（Terran）、虫族（Zerg）、神族（Protoss）。每个种族均包括多种生命角色、战斗装备、功能建筑等多类型单元。三种角色各具特色：

图1　星际争霸I竞赛环境

Fig.1　StarCraft I competition environment

图2　星际争霸II游戏环境

Fig.2　StarCraft II game environment

人族：人族单元灵活、多样，其平衡了虫族和神族的特点，是两者性能的均衡。其作战单元和建筑有陆战队员、攻城坦克、巡洋舰、导弹发射塔等。

虫族：虫族繁衍迅速，需要的资源少，单位能力弱但速度快，常以成群的形式以数量占据对抗优势。其作战单元和建筑有小狗、蟑螂、飞龙、孢子塔等。

神族：神族繁殖率不高，但单元科技水平很高、能力强，因此需要的资源也多，常以策略的质量取代数量占据对抗优势。其作战单元和建筑有狂热者、圣堂武士、凤凰战机、光子炮等。

在多人对抗模式中，玩家需要收集尽可能多的矿物、天然气或零散的奖励等资源来建造更多的生产、防御等建筑物和生产更多的作战单元并提升建筑单元和作战单元的技能等级，以最短的时间消灭敌方来赢得胜利。

2.2　星际争霸研究的难点及其对人工智能研究的挑战

与棋类游戏相比，多智能体实时策略游戏相关研究更难，主要体现在以下几点。

（1）多玩家共存、多异构智能体合作。与棋类游戏博弈双方交替进行动作不同，实时策略游戏中多玩家同时推动游戏情节发展，不同的玩家可以同时进行动作。游戏中有不同的角色单元和功能建筑，如何更好地发挥每个单元的功能也是需要考虑的问题。

（2）实时对抗及动作持续性。实时策略游戏是“实时”的，意味着玩家需要在很短的时间内进行决策并行动。与棋类游戏中玩家有几分钟的决策时间不同，星际争霸游戏环境以24帧/秒频率改变，意味着玩家可以以最高不到42毫秒的频率进行动作。若以环境改变每8帧玩家进行一个动作的平均水平来看，玩家仍需要以每秒3个动作的频率进行博弈。不仅如此，玩家输出的动作有一定的持续性，需要在一定的时间持续执行，而非棋类游戏玩家的动作是间断的、突发的、瞬时的。

（3）非完整信息博弈和强不确定性。多数实时策略游戏是部分可观测的，玩家仅能观察到自己已经探索的部分地图情况。在星际争霸中，因为有战争迷雾的存在，玩家只能看到自己所控制的游戏角色当前所处环境的情况，其它环境信息无法获知。而棋类游戏玩家可以获取全棋盘的情况。多数实时策略游戏具有不确定性，即决策过程中采取的动作都有一定概率促成最后的胜利。

（4）巨大的搜索空间及多复杂任务。实时策略游戏更复杂，其在状态空间的规模上和每个决策环节可选择的动作序列均非常巨大。例如，就状态空间而言，一般的棋类游戏状态空间在10^{50}左右，德州扑克约为10^{80}，围棋的状态空间为10^{170}。而星际争霸一个典型地图上的状态空间比所有这些棋类的状态空间都要大几个量级。以一个典型的128 × 128像素地图为例，在任何时候，地图上可能会有5~400个单元，每个单元都可能存在一

个复杂的内在状态（剩余的能量和击打值、待输出动作等），这些因素将导致可能的状态极其庞大。即便是仅仅考虑每个单元在该地图上可能的位置，400个单元即有$(128\times128)^{400}=16384^{400}\approx10^{1685}$种可能。另一种计算复杂度的方式以$b^d$来计算游戏的复杂度，其中国际象棋$b\approx35$，$d\approx80$，围棋$b\approx30\sim300$，$d\approx150\sim200$，而星际争霸$b$的范围是$10^{50}\sim10^{200}$，$d\approx36000$。

多智能体实时策略游戏的这些突出难点给该领域人工智能研究方法带来巨大挑战。文献［2］将本领域研究中的挑战总结为规划、学习、不确定性、时空推理、领域知识开发和任务分解六个方面。在此基础上，我们将当前研究中的挑战分为多尺度规划与多层次决策一致性、多途径策略学习、降低不确定性、空间和时间上的多模联合推理、领域知识开发和多层次任务分解六大挑战。本领域研究难点与研究挑战的对应关系如图3所示。

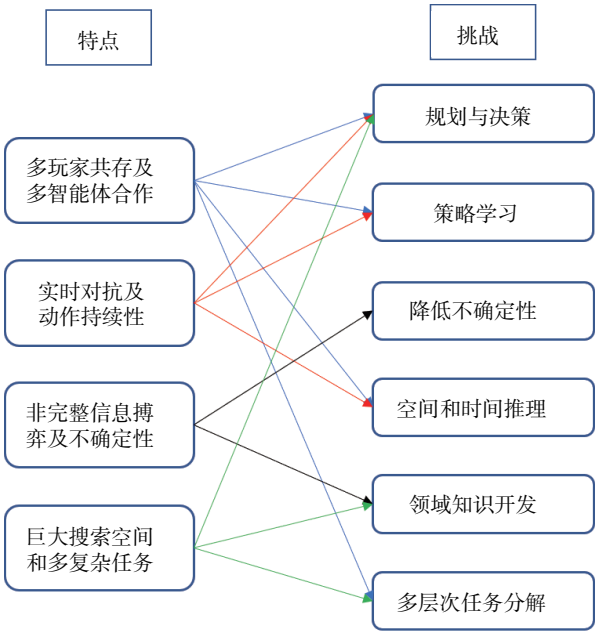

图3　多智能体实时策略游戏存在的难点与人工智能研究挑战的对应关系

Fig.3　Correspondence between the difficulties of multi-agent real-time strategy games and the challenges of artificial intelligence research

（1）多尺度规划与多层次决策一致性。一方面，由于多智能体游戏中巨大的状态空间和可输出动作，使得一般的对抗规划方法如博弈树搜索已不能满足需求，多智能体实时策略游戏需要多尺度的规划。另一方面，实时约束为多异构智能体大量的低层次动作规划与高层次全局决策目标的一致性耦合带来很大困难，难点在于设计一种既考虑复杂多目标优化又兼顾计算效率的方法，最终形成多智能体整体行动的实时一致性。

（2）多途径策略学习。除对抗规划技术之外，一些研究团队将注意力放在多途径策略学习技术上，其中包含三种策略学习问题。一是提前学习，即开发已有数据，如已有游戏回放、已有的针对特定地图的适当策略等。难点在于策略的抽象表达方法以及在实际博弈过程中如何合理选择并应用这些策略。另外，这些针对特定环境的策略是否具有普适性也有待验证。二是游戏中学习，即在博弈过程中在线学习提升游戏水平，这些技术涉及到强化学习方法及对手建模等，其难点在于状态空间巨大且部分可观测。三是游戏间相互学习，即如何将从一个游戏中学到的知识用在另一个游戏中以提升胜率。一些工作是利用简单博弈论方法从预先定义的策略池中挑选合适的策略，但这些固定的策略无法根据具体对抗环境进行自适应调整和策略提升，因此也限制了对抗的竞技水平。

（3）降低不确定性。这里的不确定性主要包括两个部分。一是由于游戏是部分可观测的，玩家无法看到全局的情况，因此需要去侦察来了解更多的情况。难点在于如何设计具有自适应能力的好的侦察策略和知识表示来降低不确定性。除此之外，由于敌人的策略也是未知的，这种不确定性造成决策的无目的性，不能很好地根据敌人的策略适时调整对抗策略，所以需要通过好的预测模型预测对手的意图。

（4）空间和时间上的多模联合推理。空间上的推理包括不同功能建筑建造的位置、防御攻击建筑建造的位置以及对战中各作战单元所处的位置等应该如何合理安排。除此之外，各功能单元在不同的地形上可以发挥出不同程度的攻击、防御

等功能，如坦克在高地势时攻击范围更大等，这些也是空间推理应考虑的因素。时间推理是指玩家既要在当前战斗中采取战术战胜敌人，又需要在更高水平上长远地规划如何安排自己的资源、建造功能建筑或升级、策略转换等。有些策略是短时间就可以看到效果的，而有些策略需要较长的时间才发挥作用，因此需要长远全局规划和短期局部规划的统一。其中长远策略规划中由于一些策略在很长一段时间后才发挥作用，导致智能体在学习过程中不能很好地从长时间的延迟奖励中学到有用的策略。另外，由于空间推理和时间推理是两种不同模式的推理形式，需要构建两种模式相融合的推理策略。

（5）领域知识开发。实时策略游戏已经发展了多年，产生很多可利用的战术动作、规律和策略等数据。充分利用该领域的已有知识可极大提升自主游戏程序的竞技水平。该领域早期研究者将从数据中总结的策略编写成代码，游戏程序可以从这些编好的代码中选择。近两年大量的游戏数据集可供机器学习提取有用信息。如何从大量的数据中提取有价值的策略，形成自主游戏程序的决策网络，仍存在极大挑战。

（6）多层次任务分解。多层次任务分解是指将多智能体博弈游戏分解成不同的子任务，通过分别解决这些子任务来降低整体解决的难度。主要可分解成以下几部分：策略，即高水平决策，如全局战役主要用什么策略；战术，即当前策略、短时策略，如一场战斗中采取何种策略；反应控制，即战斗、战术实施，如战斗中应采取何种走位、用哪种武器攻击等；地形分析，主要包括敌我双方所处位置、战斗地形、可通过道路、地势等信息；智能收集信息，主要包括敌方建造了何种建筑、生产了哪种类型的战斗单元、正在采取什么样的策略等信息。对比而言，人类玩家在玩星际争霸时，决策常分为微观操作和宏观大规模操作。人们不需要进行复杂的多层次任务分解，只需根据具体游戏环境进行微观或宏观操作即可。

3 相关研究和成果

人工智能和游戏的研究历史可以追溯到1950年[3]。自1997年5月“深蓝”击败国际象棋大师卡斯帕罗夫起至今，已有大量的游戏程序战胜了经典游戏中的世界冠军，如跳棋、奥赛罗和拼字游戏。一些布署深度神经网络的“大脑”，甚至在极其复杂的游戏中击败了世界冠军，如围棋。

从2000年左右开始，人工智能研究人员开始关注复杂的战略模拟游戏。在早期的研究中，一些人认为，智能体需要复杂的表示和推理能力才能在这些环境中胜出，而构建上述能力是具有挑战性的。研究人员通过抽象状态缩小决策搜索空间、遗传算法学习游戏规划、使用领域知识消除静态对手假设、从专家示范中提取行为知识等方法降低搜索的难度，为自主游戏程序赋予更强的能力。

在众多实时策略游戏人工智能研究环境中，星际争霸相比之前大多数工作更具挑战性。该游戏自1997年出现至今吸引了大量人类玩家，并举办了各种级别和类型的国际性赛事。2010年起，以AIIDE、SSCAIT、CIG为代表基于星际争霸I环境的各类人工智能比赛开始举办，阿尔伯塔大学、斯坦福大学、Facebook等众多高校和研究单位投入其中。这期间的人工智能算法一般被称为经典人工智能程序，大多数基于规则。这类自主游戏程序可以打败游戏内置程序，但是远远比不上人类专业选手，甚至连普通选手也打不过。

2016年开始，以深度学习和深度强化学习为主的智能体自主学习方法开始应用于该领域，此类算法被称为现代人工智能程序。Deepmind和暴雪联合开发了基于星际争霸II的深度学习研究环境SC2LE。国内外众多极具实力的科研团队参与其中，国外有如Deepmind、Facebook、阿尔伯塔大学、牛津大学、伦敦大学等，国内如阿里巴巴、腾讯以及中国科学院自动化研究所等也进行了相关研究。

2009年开始，星际争霸相关研究成果开始发

表1　星际争霸主要研究单位和方法

Table 1　The main research groups and methods of StarCraft

序号	时间	科研单位	方法
1	2009.9	美国加州大学圣克鲁兹分校	数据挖掘为对手建模[1]
2	2011.11	法兰西学院	贝叶斯模型预测策略构建树[1]
3	2012.10	美国俄勒冈州立大学	动态贝叶斯网络战略模型[1]
4	2012.11	韩国世宗大学	侦察算法和机器学习预测对手[1]
5	2012.12	新西兰奥克兰大学	一步 Q 学习和 Sarsa[2]
6	2013.8	韩国首尔大学 延世大学	用回放进行预测[1]
7	2013.12	新西兰奥克兰大学	扩展神经进化算法[1]
8	2014.3	美国东北大学	战斗近似预测模型[1]
9	2014.10	加拿大阿尔伯塔大学	逻辑回归从重放中学习模型权重[1]
10	2016.1	加拿大阿尔伯塔大学	启发式搜索和分层投资组合搜索[1]
11	2016.5	美国德雷塞尔大学	从数据回放中学习前向模型[1]
12	2016.5	美国纽约大学 Facebook 人工智能实验室	多智能体通信网络 CommNet[3]
13	2016.9	Facebook	深度神经网络结合启发式强化学习[2]
14	2017.2	英国牛津大学 微软	多智能体强化学习中重要性采样和手动调参[2]
15	2017.3	英国伦敦大学学院 阿里巴巴团队	基于 Actor-Critic 的双向协调网络 Bicnet[2]
16	2017.5	英国牛津大学 微软	基于多 Actor-Critic 的反事实多智能体（COMA）策略梯度算法[2]
17	2017.7	丹麦哥本哈根信息技术大学	深度学习回放中的宏观管理决策[3]
18	2017.8	DeepMind Blizzard	深度强化学习（A3C）[2]
19	2018.6	Deepmind	自我关注深度强化学习[2]
20	2018.9	腾讯人工智能 实验室 美国罗切斯特大学和西北大学	平层宏观动作程序（强化学习）结合分层组织的宏微观混合 动作程序[2]

注：表中上角标（1）为经典机器学习方法，（2）为强化学习方法，（3）为深度学习方法。

表。我们选出有代表性的成果进行统计（详见表1），并在下一章节中进行分类分析。

4　研究方法

本文将相关领域的研究方法分为基于规则、经典机器学习、深度学习、强化学习及其它有潜力的发展方向五类，并将指出这些方法适用于解决哪一类挑战。

4.1　基于规则

基于规则的方法用于解决策略学习和领域知识利用的挑战。这些方法将人类玩家在实践中总结出的规则编写成程序，作为自主游戏程序的一个策略模块，游戏程序在游戏进行时根据游戏的情况选择对应的策略执行即可。Certicky M[4]等根据熟练玩家用建筑物阻挡敌人进入的策略编写了自主游戏程序。提供一个准备使用的声明式解决方案，采用答案集编程（ASP）的范例，使自主游戏程序也具备合理布局建筑物来阻止敌人进入的技能。Weber B[5]等以反应性计划语言ABL构建了在游戏中指挥个体单位的游戏程序，这种反应式规划是控制低级单位命令的合适技术，部分减少了玩家需要控制的个体单位。

4.2 经典机器学习

我们将除深度学习、强化学习和深度强化学习之外的机器学习方法归为经典机器学习方法。根据各方法对应解决多尺度规划与多层次决策一致性、多途径策略学习、降低不确定性以及领域知识开发利用四类挑战，将经典机器学习方法分为快速搜索与规划、对手策略建模和作战模型、降低不确定性、行为知识提取和利用四类方法。

4.2.1 快速搜索与规划

规划与决策问题主要关注自主游戏程序不同层次的对抗策略如何优化生成。David C[6]在星际争霸人工智能竞赛中使用在线的启发式搜索算法，该搜索算法能够实时生成专业人类玩家水平的构建命令。其为考虑时长、持续时间、投资组合的贪婪搜索分别设计了三种单位微观管理算法，并将分层投资组合搜索用于搜索巨大的游戏空间。Aha D W[7]等在搜索内部空间的遗传算法以及偏向子计划检索的加权算法基础上改进，引入一个计划检索算法，消除了前两种方法假设静态对手的不足，由此可将学习的知识扩展到具有完全不同策略的对手。Zhen J S[8]等使用扩展拓扑的神经进化（NEAT）算法，以增强人工智能游戏程序的适应性，实现快速、实时评估和反应。

4.2.2 对手策略建模和作战模型

策略学习问题主要关注如何从回放数据中学到有用的知识。Weber B G[9]用数据挖掘方法从大量的游戏日志中学习高水平玩家的策略，并为游戏中的对手建模，以此在游戏中检测对手策略，预测对手什么时候执行策略并做出行动。Uriarte A[10]等从回放数据中学习作战模型并用它们来模拟实时策略游戏中的战斗。

4.2.3 降低不确定性

不确定性问题一般可由为对手建模、为游戏建模的方法来进行预测，或者使用侦察算法等获取更多的信息来降低不确定性。Gabriel S[11]等通过使用贝叶斯建模来替代布尔值逻辑，处理信息的不完整性和由此产生的不确定性。通过机器学习从高水平玩家的回放数据来对动态对手建模，进行战略和战术适应。这些基于概率的玩家模型可以通过不同的输入应用于决策，由此解决不确定情况下的多尺度决策。Park H[12]使用侦察算法和机器学习算法来预测对手的攻击时机。Hostetler J[13]等提出动态贝叶斯网络策略模型，该模型能够从现实的观察中推断游戏的未观察部分。Cho H C[14]通过预测对手的策略改变命令顺序。Erickson G[15]提出预测游戏中哪个玩家获胜的模型。Helmke I[16]等用简单的战斗近似模型预测不涉及微观管理的战斗。Uriarte A[10]等提出了双人博弈游戏的战斗模型，用来模拟游戏中的战斗，并分析如何从回放数据中学习作战模型。

4.2.4 行为知识提取和利用

领域知识开发和利用目的是更好地利用已有的策略知识和游戏数据。Mishra K[17]等提出基于案例的实时计划和执行方法。通过以个案的形式从专家示范中提取行为知识，将这些知识通过基于案例的行为生成器调用形成合适的行为，来实现当前计划中的目标。Synnaeve G[18]等主张通过人类或游戏程序玩家对录制的游戏完整状态进行探索，以发现如何推理策略。他们把军队组合起来，以此减少高斯混合程度，达到在组的水平上进行战略推理的目的。

4.3 深度学习

基于深度学习的方法用于从当前大量的高水平玩家数据中学习策略，以解决领域知识开发利用的挑战。Sukhbaatar S[19]等提出一种深度神经模型CommNet，它通过使多智能体间保持连续通信来完成合作任务。该网络模型可使智能体学习彼此沟通的能力，相对于非交互智能体产生了更好的表现。Justesen N[20]等通过深度学习直接从游戏回放中学习星际争霸中的宏观管理决策。从高水平玩家的2005个回放中提取的789571个状态动作来训练神经网络，预测下一个构建动作。通过将训练好的网络整合到一个开源的星际争霸自主游戏程序UAlbertaBot中，该系统可以显著地超越

游戏内置的自主程序，并以固定的急速策略进行对抗。

4.4 强化学习

强化学习和深度强化学习一般用于解决策略学习中的挑战。我们将使用强化学习或深度强化学习的方法按照算法内容分为Q学习及其变体、Actor-Critic结构及其变体以及分布式多智能体强化学习三类。

4.4.1 Q学习及其变体

Stefan W[21]等应用Q学习和Sarsa算法的变体，使用资格痕迹来抵消延迟奖励的问题。其设计了一个能够在复杂的环境中以无监督的方式学习的智能体，替换非自适应的、确定性的游戏人工智能程序来执行任务。针对最大化奖励或学习速度两个不同的侧重点，他们证明一步式Q学习和Sarsa在学习管理战斗单元方面是最好的。Mnih V[22]等提出深度Q网络方法，可以使用端到端的强化学习直接从高维视觉输入中学习成功的策略。该方法在Atari游戏上被证明是有效的，这为用深度强化学习解决多智能体的游戏提供了思路。Kempka M[23]等在一个三维第一人称视角环境——VizDoom中验证了视觉强化学习的可行性。在一个基本的移动及射击任务和一个更复杂的迷宫导航两种场景中，使用具有Q学习和经验回放的深度卷积神经网络，都能够训练出展现人类行为的自主游戏程序。Usunier N[24]等提出深度神经网络控制器从游戏引擎给出的原始状态特征来处理微观管理场景的方法，解决了军队成员在战斗中短期低水平的控制问题。同时提出了一个结合策略空间直接探索和反向传播的启发式强化学习算法，该算法使用确定性策略来收集学习的痕迹，这比“野兽般的探索”更为有效。

4.4.2 Actor-Critic结构及其变体

Peng P[25]等在处理星际争霸中协调多个战队作战打败敌人任务时，为了保持一个可扩展而有效的通信协议，引入了一个多主体双向协调网络——BiCNet。该网络含有一个向量化扩展的Actor-Critic公式，可以处理对战双方不同类型的任意数量的智能体的战斗。在没有任何监督如人类示范或标记数据的情况下，BiCNet可以学习各种经验丰富的游戏玩家常用的高级协调策略。Foerster J[26]等提出了一种反事实多智能体（COMA）策略梯度的多智能体Actor-Critic方法。COMA使用集中的Critic来估计Q函数，用分布式的Actor来优化智能体的策略。为了解决多智能体信用分配的挑战，其使用了一个反事实的基线，边际化一个智能体的行为，同时保持其他智能体的行为固定。在具有显著局部可观的分布式多智能体情况下，COMA方法与其它多智能体Actor-Critic方法中最先进的集中控制器最好的表现对比，发现其平均性能显著提高。Vinyals O[27]等介绍了适用于星际争霸II领域的典型深度强化学习智能体的初始基线结果。在迷你游戏中，这些智能体可以通过学习达到与新手玩家相当的游戏水平。但是，在完整游戏的训练中，这些智能体无法取得重大进展。

4.4.3 分布式多智能体强化学习

Lanctot M[28]等为解决多智能体强化学习（MARL）中使用独立强化学习（InRL）策略在训练期间可能会过拟合其他智能体策略的问题，引入了一个新的度量即联合政策关联，来量化这种影响。同时提出一种通用MARL算法，该算法基于对深度强化学习生成的策略混合的近似最佳响应以及经验博弈分析来计算策略选择的元策略。Max J[29]等在第一视角多人游戏中采用双层优化的方法。一群独立的强化学习智能体通过上千种并行游戏以团队的形式在随机产生的环境中与对手进行博弈。其中这群智能体中每个个体学习其自己的内部奖励以补充来自获胜的稀疏延迟奖励，并使用新颖的时间分层表示来选择动作，使得智能体可以在多时间尺度进行推理。

4.5 其它有潜力的方向

（1）子博弈。Brown N[30]等提出用不完美信息博弈中子博弈方法解决分布式博弈和全局目标统一

的问题。该方法可用于解决多智能体实时策略游戏中分布式局部决策与团队目标统一的问题。

（2）增量学习。Xiao C J[31]等提出的增量记忆蒙特卡洛搜索树方法，为多智能体决策系统通过不断积累来提升决策能力提供潜在的可行方向。

（3）博弈论。Fang F[32]等用博弈论系统预测可能的袭击地点，打击偷猎行为。Tuyls K[33]等让智能体在非对称博弈中找纳什均衡。基于博弈论对多智能体博弈游戏分析，或许可以从更高水平的视野找到解决办法。

5 相关资源

本章介绍与星际争霸相关的资源，包括开源研究平台、开源数据集和人工智能程序竞赛。

5.1 开源研究平台

5.1.1 完整星际争霸学习环境

（1）SC2LE。Deepmind和暴雪在2017年联合推出基于星际争霸II的人工智能学习环境SC2LE。Lanctot M[28]等描述了星际争霸II领域的观察、行动和奖励规范，并提供了一个开源的基于Python的接口来与游戏引擎进行通信。除了完整的游戏地图之外，还提供了一套迷你游戏，专注于星际争霸II游戏中的不同任务。

（2）TorchCraft。Synnaeve G[34]等开发了TorchCraft，一个通过在机器学习框架Torch中控制游戏来实现诸如“星际争霸：母巢之战”等实时策略游戏深度学习研究的库。

5.1.2 类似的AI学习环境

（1）轻量级星际争霸研究环境

ELF。Tian Y[35]等提出一个覆盖范围广、轻量级和灵活的基础强化学习研究平台——ELF。ELF包含三种游戏环境（微型实时策略、夺旗和塔防）的高度可定制的实时策略引擎。其中“微型实时策略”作为星际争霸的微型版本，捕捉了关键的游戏动态，可在笔记本电脑上以每秒40K帧速运行。该系统与现代强化学习方法结合使用时，可用6个CPU和1个GPU的计算硬件在一天时间内完成端到端的完整游戏的自主游戏程序训练。此外，该平台在环境–智能体通信拓扑、强化学习方法的选择、游戏参数的变化等方面是灵活的，并且可以迁移到现有的基于C/C++的游戏环境，如ALE。

美国纽约大学和Facebook AI Research设计了一个简单的2D游戏环境，用强化学习在该环境上布署各种神经模型，在该环境中训练的模型可直接应用于星际争霸游戏[36]。

（2）其它相似研究环境

VizDoom。VizDoom是一个以第一人称视角多人射击类3D游戏Doom为基础、可进行以像素信息为输入的强化学习方法研究平台。Kempka M[23]等在该环境中验证了视觉强化学习的可行性。在一个基本的移动及射击任务和一个更复杂的迷宫导航两种场景中，使用具有Q学习和经验回放的卷积深度神经网络，都能够训练出展现人类行为的有能力的自主游戏程序。

ALE。Naddaf Y[37]介绍了街机游戏学习环境——ALE。ALE为数百个Atari 2600游戏环境提供界面，并为评估和比较强化学习、模型学习、基于模型的规划、模仿学习、迁移学习等方法提供了一个严格的测试平台。ALE提供的评估方法可以在超过55个不同的游戏中报告验证结果。

Gym。由OpenAI开发的强化学习研究环境和工具包[38]。

Minecraft。微软开发了基于Minecraft（我的世界）游戏的人工智能研究平台[39]。

另外，还有如Deepmind的Psychalab心理学实验室开发的第一人称视角3D强化学习研究环境等。

5.2 开源数据集

5.2.1 基于星际争霸II的数据集

SC2LE。Deepmind和暴雪在推出基于星际争霸II的人工智能深度学习研究环境SC2LE的同时，对于完整的游戏地图，还提供了来自人类专业玩家的游戏回放数据集，并给出从该数据训练的神

经网络来预测游戏结果和玩家行为的初始基线结果。

MSC。中科院自动化所的张俊格等发布了基于SC2LE平台的新型数据集MSC[40]。MSC由良好设计的特征向量、预定义的高水平行动和每个匹配的最终结果组成。为便于评估和比较，他们还将MSC划分为训练、验证和测试集。除了数据集之外，他们还提出了基线模型，并提出了全局状态评估的初始基线结果，构建了命令预测。为了对星际争霸II的宏观管理进行研究，还介绍了数据集的各种下游任务和分析。

5.2.2 基于星际争霸I的数据集

Facebook的Lin Z[41]等开发了基于星际争霸I的数据集。Synnaeve G[18]等提供了包含大部分游戏状态（不仅是玩家的命令）的星际争霸游戏数据集。Alberto Uriarte开发了持续更新的基于星际争霸I的高水平玩家离线数据集。

5.3 竞赛

5.3.1 AIIDE

AAAI人工智能和互动数字娱乐会议（AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment，AIIDE）[42]由人工智能促进协会（AAAI）赞助，每年举行一次。会议展示关于娱乐中智能系统建模、开发和评估的跨学科研究，重点关注商业计算机和视频游戏。该会议长期以来一直以电脑游戏中的人工智能研究为特色，并发展到游戏以外的娱乐领域，会议上举行星际争霸人工智能自主游戏程序竞赛。会议从2005年开始，已经举办了14届。

5.3.2 CIG

IEEE计算智能与游戏大会（IEEE Conference on Computational Intelligence and Games，CIG）[43]是将计算和人工智能技术应用于游戏的年度盛会。会议的领域包括适用于各种游戏的各种计算智能和人工智能，包括棋盘游戏、视频游戏和数学游戏。于2005年开始作为研讨会，自2009年开始作为会议，每年召开一次。该会议上进行星际争霸人工智能自主游戏程序比赛。

5.3.3 SSCAIT

学生星际争霸AI锦标赛（Student Starcraft AI Tournament & Ladder）[44]是一项教育活动，于2011年首次举办，是主要面向学生（非学生也允许提交）人工智能和计算机科学的竞赛。通过使用BWAPI提交用C++或Java编程的自主游戏程序来进行一对一星际争霸游戏。

6 未来研究趋势

非完美信息下的多智能体博弈研究是当前众多人工智能研究团队努力攻克的难题，虽然有新的成果不断产生，但直到目前，完整游戏情况下，人工智能游戏程序仍无法达到人类高水平玩家的水平。为了达成这一目标，除了文章前述的研究方法之外，一些研究者将注意力放在多智能体分布式决策上。分层和分任务决策对星际争霸来说可能是一种发展方向，通过将对抗任务分不同的层次和拆分成不同的任务模块，在小的任务范围内进行学习，最终将这些模块整合成一个完整的人工智能游戏程序。另外，将博弈论作为对抗分析的指导方法，会给该领域带来新的解决思路。除此之外，模仿学习、迁移学习以及增量式学习也可能在该领域展现出好的效果。

多智能体对抗博弈策略在一些实际领域具有应用价值。其中简单任务应用如追捕任务，即多机器人协同追捕“逃跑者”机器人。与之类似，有多机器人协同阻止入侵者的“疆土防御”任务。机器人足球是更高水平的复杂任务，各足球机器人需要团队协作采取策略与对手机器人团队进行对抗，防守好自己的球门并尽可能多地进球得分。值得注意的是，当前多智能体对抗博弈策略研究在军事领域受到重点关注。以美国军方为例，其连续几年发布的无人系统路线图均将多无人系统在战场中的协作作战列为重点发展方向，并进行了多项以多机器人系统或集群作战为内容的军事研究项目。另外，俄罗斯军方已将多无人系统应

用于实际战场。

目前，多智能体博弈游戏仍是一个开放的难题，人工智能游戏程序还无法超越人类顶级玩家的水平。随着人工智能技术的快速发展以及越来越多科研团队投入其中，该领域将会有更多更震撼的成果陆续产生。

参 考 文 献

[1] 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.

[2] Ontanon 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 & Ai in Games, 2013, 5 (4): 293-311.

[3] Aha D W, Molineaux M, Ponsen M. Learning to win: case-based plan selection in a real-time strategy game [C]. International Conference, on Case-Based Reasoning, ICCBR 2005, Chicago, USA, August 23-26, 2005.

[4] Certicky M. Implementing a wall-in building placement in StarCraft with declarative programming [J]. Eprint ArXiv, 2013.

[5] Weber B. Reactive planning for micromanagement in RTS games [R]. University of California, Santa Cruz, 2014.

[6] David C. Heuristic search techniques for real-time strategy games [D]. Edmonton: University of Alberta, 2016.

[7] Aha D W, Molineaux M, Ponsen M. Learning to Win: Case-Based Plan Selection in a Real-Time Strategy Game [J]. 2005.

[8] Zhen J S, Watson I. Neuroevolution for micromanagement in the real-time strategy game Starcraft: Brood War [M]. AI 2013: Advances in Artificial Intelligence. Springer International Publishing, 2013: 259-270.

[9] Weber B G, Mateas M. A data mining approach to strategy prediction [C]. IEEE Symposium on Computational Intelligence and Games, 2009, CIG 2009, 2009: 140-147.

[10] Uriarte A, Ontanon S. Combat models for RTS games [J]. IEEE Transactions on Computational Intelligence & Ai in Games, 2016, 99: 1-1.

[11] Gabriel S. Bayesian programming and learning for multi-player video games: application to RTS AI [D]. Lyon: University of Grenoble, 2012.

[12] Park H, Cho H C, Lee K Y, et al. Prediction of early stage opponents strategy for StarCraft AI using scouting and machine learning [C]. Workshop at SIGGRAPH Asia. ACM, 2012: 7-12.

[13] Hostetler J, Dereszynski E, Dietterich T, et al. Inferring strategies from limited reconnaissance in real-time strategy games [C]. Twenty-Eighth Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2012: 367-376.

[14] Cho H C, Kim K J, Cho S B. Replay-based strategy prediction and build order adaptation for StarCraft AI bots [C]. IEEE Computational Intelligence in Games, 2013: 1-7.

[15] Erickson G, Buro M. Global state evaluation in StarCraft [C]. Tenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AAAI Press, 2014: 112-118.

[16] Helmke I, Kreymer D, Wiegand K. Approximation models of combat in StarCraft 2 [J]. Eprint Arxiv, 2014.

[17] Mishra K, Sugandh N, Ram A. Case-based planning and execution for real-time strategy games [C]. International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development. Springer-Verlag, 2007: 164-178.

[18] Synnaeve G, Bessiere P. A dataset for StarCraft AI & an example of armies clustering [J]. Eprint Arxiv, 2012.

[19] Sukhbaatar S, Szlam A, Fergus R. Learning multiagent communication with backpropagation [C]. 29th Conference on Neural Information Processing Systems (NIPS 2016), 2016.

[20] Justesen N, Risi S. Learning macromanagement in StarCraft from replays using deep learning [J]. Eprint ArXiv, 2017.

[21] Stefan W, Ian W. Applying reinforcement learning to small scale combat in the real-time strategy game StarCraft: Broodwar [C]. 2012 IEEE Conference on Computational Intelligence and Games (CIG), 2012: 402-408.

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

[23] Kempka M, Wydmuch M, Runc G, et al. ViZDoom: A

Doom-based AI research platform for visual reinforcement learning [C]. IEEE Computational Intelligence and Games, 2017: 1-8.

[24] Usunier N, Synnaeve G, Lin Z, et al. Episodic exploration for deep deterministic policies: an application to StarCraft micromanagement tasks [J]. Eprint ArXiv, 2016.

[25] Peng P, Wen Y, Yang Y, et al. Multiagent bidirectionally-coordinated nets: emergence of human-level coordination in learning to play StarCraft combat games [J]. Eprint ArXiv, 2017.

[26] Foerster J, Farquhar G, Afouras T, et al. Counterfactual multi-agent policy gradients [C]. The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), 2018.

[27] Vinyals O, Ewalds T, Bartunov S, et al. StarCraft II: a new challenge for reinforcement learning [J]. Eprint ArXiv, 2017.

[28] Lanctot M, Zambaldi V, Gruslys A, et al. A unified game-theoretic approach to multiagent reinforcement learning [C]. The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS), 2017.

[29] Max J, Wojciech M C, Iain D, et al. Human-level performance in first-person multiplayer games with population-based deep reinforcement learning [J]. Eprint ArXiv, 2018.

[30] Brown N, Sandholm T. Safe and nested subgame solving for imperfect-information games [C]. The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS), 2017.

[31] Xiao C J, Mei J C, Martin M. Memory-augmented Monte Carlo tree search [C]. The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), 2018.

[32] Fang F, Nguyen T H, Pickles R, et al. PAWS — a deployed game-theoretic application to combat poaching [J]. Ai Magazine, 2017, 38 (1): 23-36.

[33] Tuyls K, Pérolat J, Lanctot M, et al. Symmetric decomposition of asymmetric games [J]. Scientific Reports, 2018.

[34] Synnaeve G, Nardelli N, Auvolat A, et al. TorchCraft: a library for machine learning research on real-time strategy games [J]. Eprint ArXiv, 2016.

[35] Tian Y, Gong Q, Shang W, et al. ELF: an extensive, lightweight and flexible research platform for real-time strategy games [C]. The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS), 2017.

[36] Sainbayar S, Arthur S, Gabriel S, et al. MazeBase: a sandbox for learning from games [J]. Eprint ArXiv, 2015.

[37] Naddaf Y, Naddaf Y, Veness J, et al. The arcade learning environment: an evaluation platform for general agents [J]. Journal of Artificial Intelligence Research, 2013, 47 (1): 253-279.

[38] OpenAI Gym [EB/OL]. http://gym.openai.com/.

[39] Matthew J, Katja H, Tim H, et al. The malmo platform for artificial intelligence experimentation [C]. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), 2016.

[40] Wu H, Zhang J, Huang K. MSC: A dataset for macro-management in StarCraft II [J]. Eprint ArXiv, 2017.

[41] Lin Z, Gehring J, Khalidov V, et al. STARDATA: a StarCraft AI research dataset [J]. Eprint ArXiv, 2017.

[42] AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE) [EB/OL]. http://www.aaai.org/Conferences/conferences.php.

[43] IEEE Conference on Computational Intelligence and Games (CIG) [EB/OL]. https://cilab.sejong.ac.kr/sc_competition/.

[44] Student Starcraft AI Tournament & Ladder [EB/OL]. https://sscaitournament.com.

作者简介：

张宏达(1991-)，男，博士研究生，主要研究方向为移动机器人系统。

李德才(1983-)，男，博士，副研究员，主要研究方向为智能机器人。

何玉庆(1980-)，男，博士，研究员，主要研究方向为机器人自主行为方法、非线性系统估计与控制、移动机器人系统。