基于深度强化学习的对手建模方法研究综述

徐浩添，秦龙，曾俊杰，胡越，张琪*
(国防科技大学 系统工程学院，湖南 长沙 410073)

摘要：深度强化学习是一种兼具深度学习特征提取能力和强化学习序列决策能力的智能体建模方法，能够弥补传统对手建模方法存在的非平稳性适应差、特征选取复杂、状态空间表示能力不足等问题。将基于深度强化学习的对手建模方法分为显式建模和隐式建模两类，按照类别梳理相应的理论、模型、算法，以及适用场景；介绍基于深度强化学习的对手建模技术在不同领域的应用情况；总结亟需解决的关键问题以及发展方向，为基于深度强化学习的对手建模方法提供较全面的研究综述。

关键词：深度强化学习；对手建模；博弈论；心智理论；表征学习；元学习
中图分类号：TP391.9 文献标志码：A 文章编号：1004-731X(2023)04-0671-24
DOI: 10.16182/j.issn1004731x.joss.22-0555
引用格式: 徐浩添, 秦龙, 曾俊杰, 等. 基于深度强化学习的对手建模方法研究综述[J]. 系统仿真学报, 2023, 35(4): 671-694.
Reference format: Xu Haotian, Qin Long, Zeng Junjie, et al. Research Progress of Opponent Modeling Based on Deep Reinforcement Learning[J]. Journal of System Simulation, 2023, 35(4): 671-694.

Research Progress of Opponent Modeling Based on Deep Reinforcement Learning

Xu Haotian, *Qin Long*, *Zeng Junjie*, *Hu Yue*, *Zhang Qi**
(College of Systems Engineering, National University of Denfense Technology, Changsha 410073, China)

Abstract: Deep reinforcement learning is an agent modeling method with both deep learning feature extraction ability and reinforcement learning sequence decision-making ability, which can make up for the depleted non-stationary adaptation, complex feature selection and insufficient state-space representation ability of traditional opponent modeling. The deep reinforcement learning-based opponent modeling methods are divided into two categories, explicit modeling and implicit modeling, and the corresponding theories, models, algorithms and applicable scenarios are sorted out according to the categories. The applications of deep reinforcement learning-based opponent modeling techniques on different fields are introduced. The key problems and future development are summarized to provide a comprehensive research review for the deep reinforcement learning-based opponent modeling methods.

Keywords: deep reinforcement learning; opponent modeling; game theory; theory of mind; representation learning; meta learning

收稿日期：2022-05-25 修回日期：2022-06-26
基金项目：国家自然科学基金(61273300，62102432，62103420)；国家社科基金军事学(2020-SKJJ-C-102)；湖南省自然科学基金(2021JJ40697，2021JJ40702)
第一作者：徐浩添(1998-)，男，硕士生，研究方向为系统仿真、多智能体系统等。E-mail：xuhaotian@nudt.edu.cn
通讯作者：张琪(1988-)，男，讲师，博士，研究方向为作战仿真、智能行为建模等。E-mail：zhangqiy123@nudt.edu.cn

0 引言

如何在合作、竞争的复杂任务场景中自主决策是当前人工智能领域所要解决的关键问题。在游戏人工智能、军事仿真、自动驾驶、机器人集群控制等应用场景的多智能体系统中，智能体具有感知、记忆、规划、决策、交流、行动等许多能力，其中对其他智能体行为、意图、信念等的推理十分重要。在此过程中，智能体往往需要通过观察其他智能体，建立除自身以外的其他智能体抽象模型，推理其行为、意图、信念等要素，并用于辅助自身决策，此过程涉及到的方法被称为对手建模(opponent modeling, OM)。

对手建模不仅关注竞争场景下的敌方智能体建模，而且还考虑合作场景下的友方建模，因此，有些文献又称其为建模其他智能体。从理论上讲，完全理性的智能体能够做出当前条件下的最优策略，实现收益的最大化。然而，现实情况下的智能体通常只具有有限程度理性[1]，决策受到情绪、偏好等影响，往往以“满意”作为收益标准。此外，基于规则的智能体，如产生式规则、启发式算法等[2-4]，遵循预置规则机制，行为模式僵硬、易于预测、理性程度不足，对手建模技术使智能体能够快速适应对手的行为方式并且在对抗中利用其弱点获取更高收益，或在合作中使团队获得更大回报。现有的对手建模方法如策略重构、类型推理、意图识别、递归推理等方法[5]，具有模型可解释、认知推理层次深的特性。然而，要进一步应用于贴近现实的问题场景仍然存在动态环境适应性弱、特征选取复杂、状态空间表示能力不足、方法在规模上的可扩展性不强等诸多缺陷。

针对以上不足，研究者们将以深度Q网络(deep Q network, DQN)[6]为代表的深度强化学习算法(deep reinforcement learning, DRL)引入到对手建模领域。其中，强化学习是智能体学习如何与环境交互，达到最大化价值和最优策略的自主决策算法。深度学习则能够从高维感知数据中提取抽象特征，对复杂的价值函数和策略函数具有很强的拟合能力。DRL有机地结合了深度学习与强化学习，前者能够增强感知与表达能力，后者提供最优决策能力，使基于DRL的对手建模(DRL-OM)技术对复杂环境中其他智能体具有更好的认知能力，目前已在德州扑克[7-8]、星际争霸II[9]等多智能体问题场景取得优异的表现。

DRL-OM是DRL方法在对手建模应用中的研究分支，涉及人工智能、神经科学、认知心理学、博弈论等众多领域。有别于以往的对手建模方法[10]，DRL-OM研究涉及更复杂的应用场景、更多元的领域交叉，在问题特性、建模方式、应用场景上和传统方法具有较大差异。虽然许多现有文献[11-12]将对手建模领域的已有研究进行了汇总分类，但目前尚没有将基于DRL方法的对手建模进行系统研究的综述文章。此外，有关多智能体强化学习的综述研究[13-14]也阐述了对手建模的应用，然而它们的内容普遍较少涉及对手建模原理，也没有系统地分类和总结对手建模方法。随着DRL越来越广泛地应用在对手建模中，领域内涌现出许多崭新的理论和方法，远超现有文献总结的涵盖范围。因此，本文将DRL算法作为研究出发点，基于对手的理性程度和建模机理提出不同于现有文献[11-12]的对手建模分类标准。此外，对手建模技术的更新迭代为现实应用提供了机遇和挑战，为此，本文汇总归纳了DRL-OM方法在应用领域的相关研究工作。

1 基本原理

1.1 对手建模原理

根据文献[5]中的定义，对手建模是一种在无法观察到其他智能体内部状态的情况下，自身智能体以观测和先验知识为输入，通过交互、观测、推理等过程捕获无法直接感知的隐藏信息，将其他一个或多个智能体的类别、目标、动作等预测

信息和意图、计划、策略等认知信息作为输出的建模方法，其核心是帮助建模者利用对手策略或根据对手隐藏的信息制定更好的响应策略。

如图1所示，对手建模可以划分为数据获取、特征提取、学习规划、策略响应4个阶段[15]。根据对手建模的应用目的，对手建模能在对抗场景中预测对手行为并予以回击，或利用对手策略上的疏漏获取更高收益。对手建模也可以在合作场景中建立团队成员的模型，在行动前予以配合，实现更为高效的任务协同。

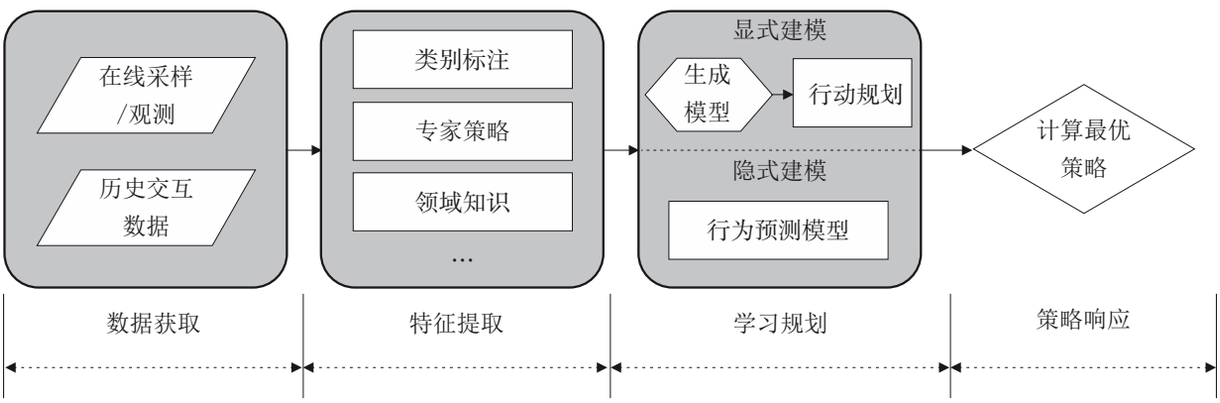

图1　对手建模方法流程
Fig. 1　Workflow of opponent modeling methods

一般来说，对手建模方法可以分为显式对手建模和隐式对手建模。前者需要建立预测对手的行为模型或者推理过程。具体来说，显式建模从与对手的在线交互中获取数据，提取对手的行动、意图、信念特征，并且将对手特征预测和规划推理两部分解耦，便于直观分析。典型的显式对手建模方法有策略重构[16-17]、递归推理[18-19]、图模型[20-21]、群体建模[22-23]等。显式建模的优点是结果预测精准，模型可信性强。缺点是特征表示复杂，需要较强的先验知识用于设计模型架构。

与显式建模方法相对，隐式建模不是建立其他智能体的显式模型，而是在自身结构或推理过程中含蓄地编码对手智能体，例如，将建模和规划合为一体或者将捕获的对手行为特征蕴含在机器学习模型的优化目标中。隐式建模方法可以通过对手交互的历史数据离线训练，避免了在线计算资源开销大、决策时延长的问题。典型的方法包括特征聚类[24-25]、神经网络预测[26]、意图识别[27-29]、类别推理[30-31]等。隐式对手建模避免了在线计算和先验知识要求，有利于在信息不完全的情况下决策。缺点是由于无法有效衡量学到的行为预测模型和真实之间的差异，导致难以权衡模型的探索和利用问题。

1.2 对手建模中的DRL算法

强化学习通常将问题建模成马尔科夫决策过程(Markov decision process, MDP)，MDP模型中状态具有马尔科夫性，即智能体的当前状态只与前一状态有关。MDP可以表示为由状态、动作、奖励、转移、折扣因子$<S, A, R, T, \gamma>$组成的五元组。针对规模较小的离散状态–动作空间，传统的强化学习方法常使用状态价值表或线性函数近似的方式求解。

在强化学习方法的基础上，DRL将深度神经网络(deep neural network, DNN)作为价值和策略的近似函数，提升了高维输入数据下的特征提取能力。DRL方法主要分为两大类[32]：值函数近似方法和策略梯度方法。值函数近似方法通过神经网络梯度下降逼近动作价值函数，缺点是容易收敛到局部最优点，难以处理连续动作空间的任务；策略梯度方法是将策略参数化，通过深度神经网络逼近策略函数，再沿着策略梯度方向求解最优策略。

1.2.1 值函数近似方法

Q学习算法是无模型强化学习的一种，通过单步时序差分学习最优行动值。Q学习算法的更新公式为

$$Q^*(s,a)\leftarrow Q(s,a)+\alpha(r+\gamma\max_{a'}Q(s',a')-Q(s,a)) \tag{1}$$

式中：(s,a)为当前状态和采取的动作；(s',a')为下一状态和采取的下一动作；α为动作价值函数的学习率；r为当前状态返回的奖励；γ为衰减因子。动作价值函数$Q(s,a)$通过式(1)时序差分近似到最优动作价值函数Q^*。

DQN将式(1)用于构造卷积神经网络的损失函数，输入当前状态信息，利用Q学习计算所有动作价值并选出最大值的动作输出。此外，DQN采用离线策略更新提高样本利用率，通过使用经验回放池减少样本之间的时序关系，提高了深度神经网络训练的收敛性，并且在预测网络更新时，固定目标网络参数，提高了Q学习的稳定性。

DQN存在诸多不足，如经验回放池容量有限、需要完整观测信息。深度循环Q网络(deep recurrent Q network, DRQN)[33]使用长短时记忆网络代替全连接层网络，用于解决部分可观测条件下的决策问题。

1.2.2 策略梯度方法

DQN无法处理连续动作控制任务，因此Lillicrap等[34]在确定性梯度理论基础上结合Actor-Critic框架提出了深度确定性策略梯度算法(deep deterministic policy gradient algorithms, DDPG)。DDPG将动作策略和动作策略探索的学习更新分离，执行动作策略时使用确定性策略，探索时使用随机策略。类似于DQN，DDPG训练时从经验回放池中采样，对目标网络和预测网络进行离线策略更新。每个网络采用Actor-Critic结构，分别进行策略更新和值函数更新，使Q网络更容易收敛。

根据异步强化学习的思想，Mnih等[35]提出异步优势演员–评论家算法(asynchronous advantage actor-critic, A3C)。A3C创建多线程同步训练，演员–评论家框架的智能体在相互间不干扰的并行环境中更新参数。A3C是多个worker从环境副本中收集经验，计算梯度之后，将梯度上传给主网络，更新主网络，并下载最新的主网络副本。避免了单一智能体连续提交更新的问题，以此降低样本相关性，提高收敛能力。

策略梯度方法具有选择更新步长的问题。步长过短会增加迭代耗时，而过长会导致更新到差策略，差策略产生的数据导致策略进一步变差，最终模型效果不佳。Schulman等[36]提出信任域策略优化算法，在策略更新中选择合适的步长，保证奖励函数单调增加。近端策略优化(proximal policy optimization, PPO)[37]在信任域策略优化算法的基础上使用一阶优化代替强约束，降低策略更新前后的分布差异性，解决了一般信任域优化算法运算复杂问题。

DRL方法在单智能体场景下大获成功，并且在多智能体场景中应用越来越广泛。对手建模技术与DRL的结合为解决多智能体决策问题带来新的突破点。表1总结了两类代表性的DRL方法，并列举了其中在对手建模领域应用较多的算法。

表1 对手建模使用的深度强化学习算法

Table 1 Deep reinforcement learning algorithms in OM

分类	算法	优点	不足
值函数近似	DQN[6]	经验复用、离轨策略机制	无法用于高维、连续空间
	DRQN[33]	采用LSTM代替全连接层	完全可观测下表现不如DQN
策略梯度	DDPG[34]	确定性策略、Actor-Critic框架	无法处理离散问题、难以确定更新步长
	A3C[35]	多线程学习、异步更新参数	更新策略方差较大
	PPO[37]	有裁剪的自适应超参数KL散度	对差异性较大样本敏感

2 基于DRL的显式对手建模

按照对手的理性程度和建模机理，基于DRL的显式对手建模的主要方法可以分为博弈均衡策略建模方法和递归推理建模方法。博弈均衡策略方法主张对手的绝对理性，采用博弈论的均衡原理建模，递归推理方法强调了智能体具备有限理性，从认知心理学的人类理性研究中获取灵感。

基于博弈均衡策略的建模方法尝试通过智能体自演收敛至博弈均衡，并且在序列、非零和等类型的博弈中追求绝对理性。主要方法有：①基于虚拟自博弈的建模方法；②基于反事实遗憾值最小化的建模方法；③基于MiniMax均衡的建模方法。

2.1 博弈均衡策略建模

2.1.1 博弈论基础概念

博弈论与多智能体系统的结合由来已久，许多对手建模方法受到博弈论的启发。在这类方法中，DRL利用强化学习的状态采样更新代替了传统的状态空间遍历，提升了学习最优策略的效率，采用深度学习的非线性近似能力来拟合更高维的博弈问题。

纳什均衡是博弈论中标志性的均衡解，代表着任何人无法通过单独偏离纳什均衡策略而获得更高收益。

定义1　纳什均衡(Nash equilibrium, NE)以玩家同时决策的策略型博弈为例，将其形式化为集合$G=(N, A_i, f_i), i\in N$，其中，N为玩家集合；A_i为玩家i的策略集合；f_i为玩家i的收益函数。博弈G的纳什均衡满足

$$f_i(s_i^*, s_{-i}^*)\geqslant f_i(A_i, s_{-i}^*),\ \forall i\in N \tag{2}$$

式中：s_i^*为玩家i采取的最优策略。

定义2　虚拟博弈(fictitious play, FP)是一类求解纳什均衡的方法，它根据对手的平均经验做出最优策略，将最优策略加入平均经验，并以此反复迭代的学习过程。如公式(3)所示，第n次迭代的智能体平均经验为σ_n，最优策略为$b(\sigma_n)$。

$$\sigma_{n+1}\in(1-1/(n+1))\sigma_n+1/(n+1)\cdot b\sigma_n \tag{3}$$

定义3　扩展式博弈(extensive-form game, XFG)可表示为有序的向量$\boldsymbol{\Gamma}=(N, V, E, x^0, V_i, O, u), i\in N$，其中，$N$为参与人的有限集合；$(V, E, x^0)$是博弈树，$V_i$为非叶节点；$O$为博弈可能结果集合；$u$为从博弈终点至集合$O$的效用映射。

博弈树搜索是另一类求解纳什均衡的学习方法，包括反事实遗憾值最小化(CFR Minimization, counterfactual regret minimization)和蒙特卡罗树搜索等，它们通过遍历博弈树计算扩展型博弈的状态、动作价值。

定义4　MiniMax均衡是博弈中具有安全性象征的均衡解，代表着玩家选择获得最高收益的保底策略。对于策略型博弈G，玩家i的最大最小值满足

$$f_{-i}=\max_{s_i\in A_i}\min_{s_{-i}\in A_{-i}} f_i(s_i, s_{-i}) \tag{4}$$

2.1.2 虚拟自博弈

虚拟自博弈[38]的主要思想是在扩展式博弈中跟踪对手的历史行为，并根据对手的平均策略选择最佳对策。神经虚拟自博弈(neural fictitious self-play, NFSP)[39-41]以虚拟自博弈为基础，利用深度神经网络拓展了不完全信息博弈的研究，并在诸如扑克的场景下达到近似纳什均衡。如图2所示，NFSP的最佳反应依赖于最佳响应网络和历史平均网络。最佳响应网络的实现是基于DQN，以$\epsilon-$贪婪的策略探索行为奖励，学习Q值得到对其他智能体历史行为的最佳响应。历史平均网络是模仿对手智能体历史最佳响应的对手模型，使用多层神经网络通过监督学习实现对从状态到行为的对手策略映射。在不完全信息博弈的二人德州扑克游戏中，NFSP智能体能够在没有领域先验知识的情况下收敛至近似纳什均衡。然而基于DQN的最优响应在对手策略变化的博弈中难以收敛，并且面临在搜索规模巨大、搜索深度较深的场景学习困难等问题。针对NFSP无法收敛最优的问题，

Zhang 等[41]提出将蒙特卡罗树与 NFSP 相结合，以在线更新策略的蒙特卡罗搜索的方式训练，解决了 DQN 无法近似最优解的问题，缺点是蒙特卡罗搜索存在产生样本方差大的问题。

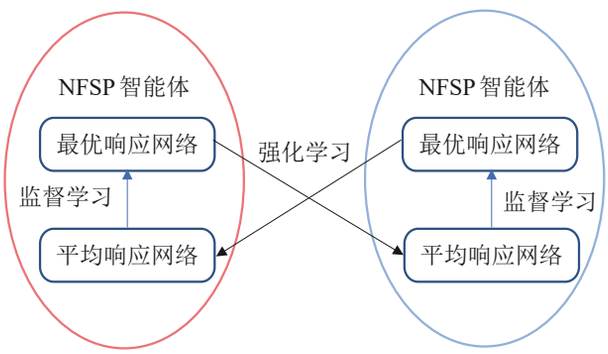

图2　NFSP 智能体架构
Fig. 2　Architecture of NFSP agents

(policy-space response oracles, PSRO)[42]是在 NFSP 的概念上进一步推广，证明虚拟博弈是以先前近似最佳反应为依据的一个特定元策略分布。假定元策略并非唯一，PSRO 通过基于线性规划、虚拟博弈、遗憾值最小化等算法实现的元策略求解器为每个智能体依次选择混合元策略。当每个智能体计算当前的元策略时，同时基于对手智能体的元策略集进行概率抽样，获得对手的混合策略，并据此计算最优响应策略更新元策略集。这类方法防止对特定策略的响应过度拟合，从而提供了一种对手规范化的形式。由于 PSRO 算法与行为博弈论模型的相似性，该算法又产生了一个独立的近似扩展版本，计算智能体的最佳响应，称为深度认知层次。在多玩家博弈中，PSRO 计算纳什均衡常面临均衡选择问题。α-rank[43]是一种用于大规模多智能体交互中评估和排序智能体的进化动力学方法。Muller 等[44-45]将 α-rank 方法应用在 PSRO 算法中，结合卡尔科夫–康尼链作为评价解概念。马尔科夫–康尼链刻画智能体种群的长期动态，作为评估对手优劣势的标准，使一般和多玩家博弈中包括非纳什均衡策略的所有策略得到充分训练改进。

2.1.3 反事实遗憾值最小化

虚拟自博弈的方法需要在真实对手交互中学习策略，然而许多应用场景无法满足与对手多次交互训练的前提条件。研究人员利用模拟的方法解决该问题，Martin 等[46]提出的反事实遗憾值最小化是利用自我博弈来最小化反事实遗憾从而拟合纳什均衡的方法，能够很好地模拟扩展式博弈的树分支节点，并在扑克游戏中取得成功。它的流程是对所有信息集 $I \in L_i$ 进行遗憾匹配计算，更新平均策略集 $\bar{\sigma}_i^t$ 并计算玩家 i 的反事实收益值 u_i^t，最后使用反事实收益更新累计遗憾值 R_i^t。然而记录所有节点的累计遗憾值导致算法的计算和存储开销巨大。遗憾匹配计算式为

$$\sigma^t(a)=\begin{cases}\dfrac{R^{t-1,+}(a)}{\sum\limits_{a'\in A}R^{t-1,+}(a')}, \text{ if } \sum\limits_{a'\in A}R^{t-1,+}(a')>0\\ \dfrac{1}{|A|}, \text{ otherwise}\end{cases} \tag{5}$$

为解决最小化反事实遗憾值方法计算速度问题，研究者们从其原理出发提出一系列在线学习方法[47-51]。Lanctot 等[48]采用蒙特卡罗采样方式对博弈树进行部分抽样，相较原方法计算效率提升 10 倍，并且证明抽样更新的遗憾收益值是真实值的无偏估计。Tammelin[51]将平均策略的收益加以衰减，提高了近期迭代策略的权重。针对 CFR 算法的存储效率问题，研究者们也提出状态空间压缩和问题抽象方法[52-53]。研究者也尝试对 CFR 方法的时效性问题和简化问题规模提出解法。王鹏程[54]提出适用于强化学习的模式匹配扑克建模方法，基于蒙特卡罗博弈树搜索计算回报函数，利用基于长短时记忆网络的 DQN 学习特征，降低了博弈问题求解难度。在非完全信息条件下，基于 CFR 方法的智能体在决策时建模对手可能采取的所有行动，将其表示为一个信息集，并学习使信息集遗憾值最小的自身策略。

2.1.4 MiniMax 均衡

上述两类模型是从纳什均衡角度求解对手建模

的最优策略，而MiniMax均衡是以安全性为前提的均衡策略对手建模方法。Wright等[55]改进了定量认知层次模型中的0级(Level-0)模型，首先手工筛选出最大最小收益、最小最大悔度等特征，将这些特征线性加权组合成为0级模型以表示智能体的初始行为，再将0级模型经过定量认知推理后得到一个有限层次推理的对手行为模型。该方法在多种游戏的玩家数据集中实验，测试结果表明对手建模能够模拟出有限程度理性的人类行为。

Li等[56]将最大最小思想作为多智能体环境中鲁棒学习的一种方法，即使智能体在训练中没有获得最优策略，学习到的鲁棒策略也使任务表现足够良好。徐浩添将MADDPG算法[57]扩展到最大最小多智能体深度确定性策略梯度(minimax multi-agent deep deterministic policy gradient, M3DDPG)[56]，该算法在考虑最坏情况的假设下更新策略，即对手模型假设所有其他智能体都将采取敌对行动，并且选择能够使自身收益最小的策略。最大最小策略的学习目标在连续动作空间计算上难以直接优化，因此，他们通过鲁棒强化学习中的最差噪声概念隐含地学习最大最小值，提出了多智能体对抗学习来解决计算最大最小策略的问题。

2.1.5 博弈均衡策略建模方法总结

基于博弈均衡策略的DRL-OM建模方法采用虚拟博弈、遗憾值最小化等学习方法，并且引入纳什均衡、最大最小均衡等均衡解概念作为效果的衡量标准。该类方法的优点是在给定适当的约束条件和博弈场景下，模型具有较强的理论支撑，以及在应对绝对理性对手时最优策略的收敛性有保证。然而，缺点是没有刻画对手特有行为习惯的能力，无法在对手模型弱点上加以利用，获得超越均衡解的收益。此外，在解决如非零和博弈、大规模零和博弈等NP−完全复杂度问题时，现有博弈论方法难以精确求解纳什均衡。表2对比了博弈均衡策略的代表性算法，分别对其研究动机、模型效果总结阐述。

表2　博弈均衡策略方法的研究动机、求解问题与效果

Table 2　Research motivation, solved problem and effects of game equilibrium strategy methods

分类	方法	研究动机	对手模型	模型效果
虚拟自博弈	FSP[38]	将FP推广至扩展式博弈	对手的历史平均最佳响应	强化学习实现最优响应，监督学习实现平均策略，收敛至纳什均衡
	NFSP[39]	使用神经网络近似最优策略和平均策略	多层神经网络近似的对手历史平均最佳响应	基于DQN实现端到端学习，并收敛至纳什均衡
	PSRO[42]	求解子博弈元策略，合并成完整策略	将博弈对手的历史策略记录在元策略集	使用DO算法[58]训练新策略，收敛性受到对手策略采样方式的影响
	α-PSRO[44]	训练改进群体的每种策略，而非单纯训练纳什均衡策略	马尔科夫–康尼链评价对手种群的质量	策略收敛于α-rank解[43]，改进了群体博弈的均衡收敛性
反事实遗憾值最小化	MCCFR[48]	采用蒙特卡罗抽样代替树节点遍历计算各个状态的遗憾值	包含对手所有可能行动的信息集	蒙特卡罗抽样是对遗憾值无偏估计，且在不完美信息扩展式博弈中快速收敛
	CFR+[51]	采用保证动作的遗憾值为正数，累计值不减少的遗憾值匹配方法	包含对手所有可能行动的信息集	改进遗憾值匹配机制，使CFR算法加速收敛近似纳什均衡
MiniMax均衡	Level-0[55]	有限理性的对手行动源自0级策略的递归推理，0级策略采用人工筛选策略	以MiniMax策略为0级的定量认知层次策略	0级策略改进认知层次模型的效果，数据集实验结果有效预测人类行为
	M3DDPG[56]	采用保底策略鲁棒应对变化对手的多智能体DRL算法	导致自身收益最小的对手策略	采用对抗学习方法求解连续动态环境的MiniMax均衡策略

2.2 递归推理建模

基于递归推理方法从认知心理学角度刻画了智能体和其他个体的心理活动，试图模仿人类社会的信念推理、意图推理等认知行为过程，期望实现智能体的有限理性决策。主要方法有：①基于心智理论的建模方法；②基于认知层次结构的建模方法；③基于贝叶斯策略复用的建模方法。

2.2.1 心智理论

心智理论是递归推理中的一类方法[59]，被认为是有助于增强深度强化学习可解释性的理论之一。在心智理论方法中，智能体对其他智能体的心理状态有明确的信念。换而言之，其他智能体的心理状态也可能包含该智能体的信念和心理状态。Rabinowitz等[60]根据心智理论提出预测对手特征和精神状态的心智理论网络(theory of mind network, ToMnet)。ToMnet通过观察获取对手的先验知识，通过元学习建立智能体的模型。ToMnet遵循一个前提假设：当遇到一个新对手时，智能体应该已经对其行为有了强大且丰富的先验知识。

如图3所示，ToMnet架构由3类网络组成：①个性网络从一组部分可观测马尔科夫过程中解析智能体的历史轨迹，形成个性嵌入e_{char}；②精神状态网络输入字符嵌入e_{char}和最近回合的轨迹输出精神状态嵌入e_{mental}；③预测网络将当前状态以及其他网络的输出作为输入，输出预测对手下步动作概率$\hat{\pi}$，对象损耗概率$\hat{c}$和后继表征$\widehat{\text{SR}}$。ToMnet正是基于3类网络的表征信息建立对手模型，在实验中分别利用贝叶斯层次推理、逆强化学习、表征学习等方法对随机策略智能体的行为进行预测；对基于目标智能体的目标进行推理；对不同种类的DRL智能体进行聚类区分；对智能体的错误信念进行判断；预测智能体的部分观察和信念。ToMnet的一个主要优点是普适性，它作为一种元学习框架，对贝叶斯最优响应、强化学习值迭代、深度强化学习网络的智能体实现方法均有出色表现。在离散网格环境中的测试结果表明ToMnet的用途广泛，能够预测通用智能体的一般行为，或者针对特定智能体预测特定的行为。同时，它的模型符合认知科学规律，具有很好的人工智能可解释性。

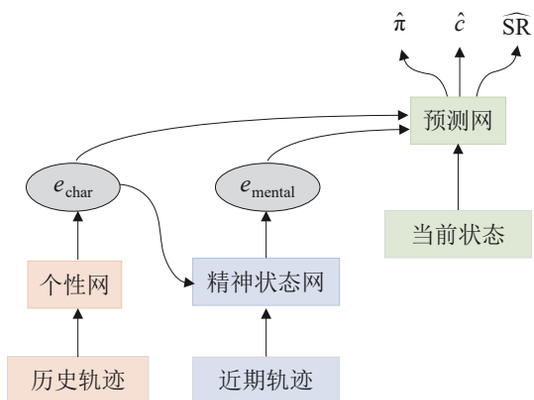

图3 心智网络架构
Fig. 3 Theory of mind network architecture

2.2.2 认知层次结构

心理学学者认为人类会将自身信念、意图和情绪归因于他人，并且使用推理能力递归地思考他人的想法，称为信念嵌套[61]。通过学习其他玩家信念状态的模型进行递归推理，这种信念的嵌套可以表现为“我相信你相信我相信”。量化认知层次结构利用量化响应和有限迭代策略推理反映有限理性的人类行为。假定每个智能体具有有限K级推理深度，Hartford等[62]将正则博弈的收益矩阵编码，构造模拟递归推理的神经网络结构。第1层网络结构是收益矩阵映射到行动向量的对手策略，作为对手模型的0级策略。从0级到K级推理的过程，前几级的策略将作为隐藏向量输入的神经网络，学习一个从隐藏层映射到动作维度的输出函数，用于模拟对手模型当前一级的策略，实现了基于认知层次结构的对手建模。该方法的优点是避免了手动筛选特征，调试了模型参数，但是该方法的架构设计较为初步，只适用于双人游戏且无重复的博弈。

Wen等[63]提出概率递归推理(probabilistic recursive reasoning, PR2)框架，考虑对手下一步行动对自身策略可能带来的影响，将对手建模转化

为推理对手下一步行动的条件概率，采用变分贝叶斯方法逼近对手的条件策略，使每个智能体找到最优策略，然后改进当前策略，在DQN和DDPG上实现自博弈收敛至纳什均衡。

然而PR2的推理层次只有1级，无法满足有限理性建模的需要。因此，为解决深层次推理的问题，Wen等[64]又在概率图模型基础上设计了广义递归推理(generalized recursive reasoning, GR2)框架，将对手下一步行动的推理拓展到了K层级推理，其中第k级推理为对1~k−1级策略联合条件概率的最优响应策略，而对手模型是基于前K−1级策略做出的最优策略。从当前状态s开始，智能体i的1级行为a_1^i取决于对手0级行动a_0^{-i}，k级行为a_k^i取决于对手的k−1级行动a_{k-1}^{-i}提供深层次推理结构的理论支撑。如图4所示，在此基础上推导出GR2演员−评论家算法，理论上证明了GR2条件下完美贝叶斯均衡的存在性，以及策略梯度方法在二人范式博弈上的收敛性。

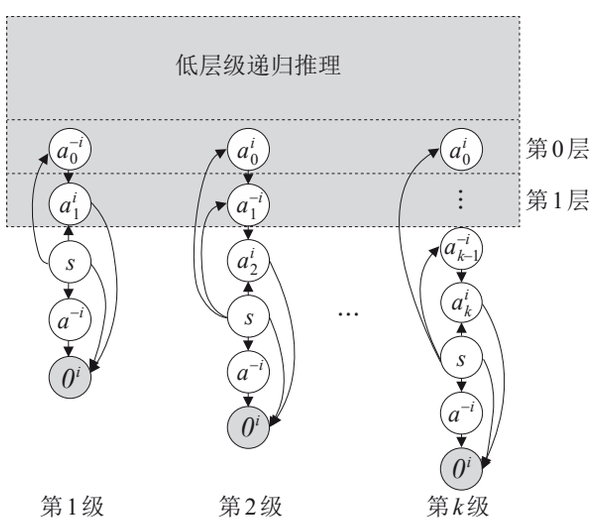

图4　递归推理过程概率图模型

Fig. 4　Probabilistic graph model of recursive inference process

2.2.3 贝叶斯策略复用

以往的对手建模通常假定智能体执行固定策略，研究者们注意到智能体可能在多种策略间切换选择，因此，设计出监测和学习对手策略的对手建模机制。切换智能体模型是一类结合贝叶斯神经网络，从观察到的状态−动作轨迹中学习对手模型的研究[65]。贝叶斯策略重用[66]是一个根据收益更新任务信念的贝叶斯模型，选择最大化收益的策略作为当前最可信的策略。

Hernandez等[67]提出改进的贝叶斯策略重用+(Bayesian policy reuse+ , BPR+)算法，将贝叶斯策略重用框架从单智能体扩展到了多智能体情景，以在线方式学习新的性能模型。BPR+方法的对手模型是一个以奖励评估对手策略信任度的贝叶斯模型，选择当前最可信的策略作为对手策略。然而BPR+的局限性是它的自我博弈行为策略表现不佳。为了提升模型的泛化性，Zheng等[68]提出蒸馏策略网络−贝叶斯策略重用+(distilled policy network-Bayesian policy reuse+, DPN-BPR+)，通过神经网络扩展了BPR+算法，并且根据行为和奖励信号共同检测对手策略。然而上述方法没有考虑具有复杂决策逻辑的对手策略识别问题，Yang等[69]从贝叶斯心智理论获得灵感，提出一种深度贝叶斯心智方法(deep Bayesian theory of mind, Deep Bayes-ToMoP)用于识别高层次推理策略，考虑策略库中对手策略的信任度以应对不同类型智能体。Deep Bayes-ToMoP根据心智理论的递归推理框架提供了一个更高层次的推理策略，识别应对未知的对手策略，面对复杂的决策逻辑也能获得智能体的最优对策。

2.2.4 递归推理建模方法总结

递归推理建模方法模拟了人类的嵌套信念推理、切换策略推理等思维逻辑，从认知机理上阐述了对手建模的过程。基于递归推理的DRL-OM方法优点是能够用来显式预测行为、意图、目标、信念等认知要素，适用于建模对手的有限理性决策过程，在人类玩家数据集和简单博弈场景得到可行性验证。缺点是目前研究进展尚为初步，现有成果只能解决少数理想条件下的问题，在复杂的实际场景尚难以应用。表3对递归推理的代表性算法研究动机、创新点与局限进行了总结阐述。

表3 递归推理方法的研究动机、创新点与局限总结

Table 3 Summary of research motivations, innovations and limitations of recursive reasoning methods

类别	算法	研究动机	模型用途	创新点	局限性
心智理论	ToMnet[60]	从心智理论提出符合人类认知的元学习对手模型	预测的对手行为、目标、信念	建立元学习的先验模型，用于预测表征和心智状态	适用的实验场景简单，环境完全可观
认知层次结构	PR2[63]	智能体具有推断对手策略的信念递归推理能力	推理对手下一步意图	提出多智能体概率递归推理的分布式框架，利用变分贝叶斯推理对手策略	二人博弈场景收敛，复杂合作场景中表现不足
	GR2[64]	借助不同层次结构的递归推理建模对手的有限理性	以K层深度推理对手的下一步意图	设计了基于概率图模型的层次结构，并证明存在完美贝叶斯均衡	具有递归推理层级选择问题，带来更高计算要求
贝叶斯策略复用	DPN-BPR+[68]	针对非平稳的对手策略，提出策略检测和复用机制	根据收益更新对当前对手策略的信念	深度神经网络作为BPR+的值函数近似，使用网络蒸馏存储最优响应策略	假定对手在固定策略之间切换，无法识别连续演化的对手策略
	Deep Bayes ToMop[69]	将BPR预测能力和心智理论的递归推理能力结合互补	在BPR信念基础上多层递归推理	具有学习对手演化和应对未知对手策略的能力	在线学习新策略的耗时长，无法应对多个对手

3 基于DRL的隐式对手建模

隐式对手建模方法没有预置形式化的对手，而是按照任务需求利用对手信息，与深度强化学习结合影响自身决策。因此，根据对手模型的作用机理，本文将基于DRL的隐式对手建模主要方法分为：①基于辅助任务的建模方法；②基于学习表征的建模方法；③基于最大化概率推理的建模方法；④基于自我–他人交互的建模方法。

3.1 辅助任务

辅助任务是一种多任务的DRL方法[70]，为智能体提供了更丰富的行为动机，适合非平稳环境的学习任务。辅助任务的建模方法设计获取对手特征的模型，修正强化学习的策略和价值函数，为强化学习提供与对手博弈的行为动机。该方法将其他智能体的观测编码作为监督信息输入神经网络进行训练，提取对手策略特征，用于修正强化学习的策略和价值函数，完成最大化奖励的目标[26,71-72]。这类方法中具有开创性的工作是He等[26]提出的深度强化对手网络(deep reinforcement opponent network, DRON)。在多智能体系统中，对手的行动将会改变环境状态转移，从而影响智能体强化学习过程的收敛性。因此，DRON同时训练Q值网络和对手策略表征网络，将对手策略表征作为环境已知条件帮助Q值网络收敛。在具体实现中，DRON在DQN的基础上改造，使用一个网络评估Q值，另一个网络则负责对手建模，以对手的行动作为输入，捕获当前对手的特征，用于学习对手策略。在此基础上，它结合多个专家网络预测估计Q值，每个专家网络捕获一种对手策略。在1v1足球比赛场景下DRON智能体与基于规则的对手智能体对抗训练，DRON智能体赢得了99.86%的进攻，并且防守成功不低于90.20%。该方法的缺点是专家网络的模型和输入需要依据先验知识手工设定。

与人工方法相比，许多研究采用监督学习的方法选取特征。例如，深度策略推理Q网络(deep policy inference Q-network, DPIQN)[71]和深度策略推理循环Q网络(deep policy inference recurrent Q-network, DPIRQN)[71]计算推断出对手策略和对手的真实观察(one-hot编码的动作向量)之间的交叉熵，提出一种自适应系数辅助修正对手模型的损失函数，用于学习对手策略的辅助任务。DPIQN和DPIRQN通过卷积神经网络处理图像数

据后，将隐藏层向量分别输入Q值网络和策略特征学习网络中。策略特征学习网络是一个多层感知机或LSTM网络，用于从隐藏向量提取对手策略特征，分别用于Q学习任务和还原对手策略的辅助任务，并根据任务损失修正网络。同样在1v1足球场景中，DPIQN智能体几乎完全赢得与基于规则对手的比赛。此外，Hernandez-Leal等[72]在A3C算法的基础上实现参数共享对手建模(agent modeling by parameter sharing, AMS)。AMS-A3C算法基于一套共享的A3C神经网络结构决策和建模对手，利用修正的损失函数完成改进自身策略的强化学习任务和监督学习拟合对手行动策略的辅助任务。实验测试中，利用CNN网络从像素游戏画面提取对手的隐含特征，用于学习预测对手策略的辅助任务网络。

3.2 学习表征

学习表征的建模方法以无监督的学习形式从观测信息中提取对手的表征信息，用于增强深度学习网络的输入信息，无需数据集、奖励等先验知识，提取表征的常见结构包括图神经网络架构[73]、前馈神经网络[74]、循环神经网络[75]等。

Tacchetti等[73]提出基于图神经网络表征的关系前向模型(relational forward models, RFM)，以环境中的对手收益作为监督信号，训练图网络在输入环境描述后，准确预测出对手的未来行为和奖励，并增强优势演员–评论家方法的表现。RFM模型是由图、边缘、节点及各自属性构成的图神经模型，其中产生的中间表征可以分析多智能体系统的社会属性，如边缘属性的大小能够解释对手采取行动的幅度，增删节点后计算边际奖励来衡量对手的社会关系强度和效用。此外，图神经网络表征具有很强的关系表示能力，具有网络拓展性，能够和RNN等神经网络结构结合。

Grover等[74]从设计优化损失函数的角度提出一种基于前馈神经网络的对手表征学习框架“Emb”，设计了一种基于编码器–解码器的无监督学习结构。根据不同策略对手表征的真实性和差异性设计损失函数，从与对手交互的历史数据中学习连续表征。随后，在PPO算法的基础上实现具有表征增强的PPO-Emb方法，验证了对手表征对强化学习策略的提升效果。Emb框架的优点是采用无监督学习对手特征，样本利用高效，无需任务领域知识。

3.3 概率推理

与学习辅助任务和表征的方法不同，通过引入表示“最优性”的二元随机变量o，将强化学习定义为一个概率推理问题。在非完全信息的博弈场景中，智能体的回报$\exp(R(s_t, a_t^i, a_t^{-i}))$由自身策略和包括对手最优策略在内的条件概率$P(o_t^i=1|s_t, a_t^i, a_t^{-i})$决定。因此，如式(6)所示，每个智能体的目标是在给定对手策略下最大化奖励。

$$\max J \triangleq \ln P(o_{1:T}^i=1|o_{1:T}^{-i}=1) \tag{6}$$

Tian等[76]提出借助行动隐式通信的(policy-belief-iteration, P-BIT)多智能体团队合作方法，P-BIT方法用于团队合作中的队友建模，模型用于推理队友的私有信息，并将其作为条件最大化的条件概率下界来求解最优策略。该算法将使用期望极大(EM)算法求解决策推理问题。P-BIT在具体流程中建立信念模块$\Phi^i(x_t^{-i}|h_t^i)$，将观测动作h_t作为输入，输出队友的私有信息x_t^{-i}，策略模块$\Phi^{-i}(x_t^i|h_t^{-i})$利用私有信息推测动作分布，2个模块迭代训练。为了实现协作场景下的多智能体隐式通信，该算法为策略模块的训练设置辅助奖励，以鼓励智能体信息交换。

最大熵学习[77]常用在单智能体的策略探索中，用于保证学习策略的多样性。Zheng等[78]将多智能体RL问题建模成贝叶斯推理并提出基于最大熵目标的正则化对手模型算法(regularized opponent model with maximum entropy objective, ROMMEO)。ROMMEO将智能体所获得的回报看作学习最优策略的概率下界，而优化智能体的回报需要对手行动。因此，为了最大化智能体回报，ROMMEO将

对手建模策略和对手动作分布之间的KL散度作为惩罚项预测对手行动，从而保持当前对手模型的真实性。同时，以增加智能体策略熵的方式提升策略随机性，用于寻找自身的最优策略。

P-BIT和ROMMEO都将智能体交互显式建模为“最优性”条件概率的问题，提供了在不完全信息条件下的团队合作建模思路。P-BIT通过推理队友私有信息来优化条件概率下界，ROMMEO则是采用最大熵的方式推理最优策略的条件概率。它们的缺点是只适用于完全合作博弈场景下的理性决策，并且最优性策略的收敛性受对手模型的准确性影响。这些局限性较大地限制了上述方法的适用范围和实用性。

3.4 自我–他人交互

前文所述方法大多通过推理对手真实行动，观察得出对手模型的特征。与此不同的是，许多研究考虑从衡量智能体自身受到对手交互影响的角度建立对手模型。例如，当所有智能体具有相同的输入变量和模型结构，以及近似的任务目标时，智能体与对手行动的差别取决于目标奖励。根据这个理念，Raileanu等[79]提出了一种推断未知对手目标的建模方法，即自我–他人建模(self-other modeling, SOM)。与以往的方法不同，SOM不是预测对手策略作为手段，而是估计对手的目标。SOM智能体使用了2个决策网络，第1个用于计算智能体自身策略，第2个是学习对手策略的对手模型。并且SOM显式地建立了智能体的目标这一概念，将目标作为对手模型的参数，根据对手模型输出与观测真实对手行动反向传播调节参数得到对手目标。SOM能够在合作和竞争性的环境中推断目标，预测出对手的行为，表现优于独立决策智能体和基于DRQN智能体。SOM算法的优点是智能体以自身模型推断对手，不需要设定额外参数来建立对手模型，便于推断任意数量智能体的目标。其缺点是具有较强的假设，即假定智能体共享一组目标，每个个体在事件开始时都被分配一个目标，并且奖励结构取决于他们所分配的共同目标。此外，SOM算法每步都需要在线优化目标估计网络，决策时间长。

智能体需要考虑交互过程中的对手学习过程，而非将其看作静态环境的一部分。根据这一思想，Zhang等[80-81]提出需要在更新自身策略梯度时考虑对手学习策略的梯度更新，并将对手的学习过程考虑在自身策略优化中。假设智能体和对手的模型参数分别为θ_1和$\Delta\theta_2$，与优化$V_1(\theta_1,\theta_2)$的常规梯度优化方法不同，研究者认为具有对手学习过程的$V_1(\theta_1,\theta_2+\Delta\theta_2)$能够避免学习静态的对手策略。Zhang等[80]假定$\Delta\theta_2$对θ_1是不可微的情况下，在二人博弈实验中取得均衡。Foerster等[81]提出的LOLA(learning with opponent learning awareness)算法则在$\Delta\theta_2$对θ_1可微的假设上利用对手学习过程优化策略，提出借助对手价值函数的二阶优化项优化自身梯度。LOLA算法在囚徒博弈中实现最大双方价值的合作策略，它的缺点是无法在对手策略不可微时使用，并且忽略了对手能够察觉到LOLA对其建模并反制的情形。

3.5 隐式建模方法总结

基于DRL的隐式对手建模方法将对手观测编码后输入深度神经网络，确立学习某种特征的目标。该类方法的优点是避免显式的复杂建模过程，能够端到端地学习预测、推理。它的缺点是基于黑盒方式构建的隐式建模依赖大量交互样本学习模型，但由于隐式模型存在不确定性，因此需要平衡探索和利用之间的关系。表4详细对比了隐式建模的各类代表性算法，分别对其研究动机、创新点与局限进行了分析总结。

4 应用场景

DRL-OM方法起初应用于游戏领域的智能体建模，随着深度强化学习算法在现实场景中的落地应用，DRL-OM技术也拓展到军事仿真、公共安全等诸多领域。

表4　隐式DRL对手建模方法的研究动机、创新点与局限总结

Table 4　Summary of research motivation, innovation points and limitations of implicit based opponent modeling methods

类别	算法	研究动机	模型特点	创新点	局限性
辅助任务	DRON[26]	设计挖掘不同对手策略隐藏特征的神经网络	使用MLP处理对手行动，将表征信息用于强化学习任务	提取对手特征用于DRL算法决策	手工提取输入专家网络的特征，可采用RNN改进
	DIPQN[71]	从观测直接提取对手策略特征，训练对手建模的辅助任务	策略特征网络学习从观测提取表征，并通过行为克隆的准确性修正网络	设计了调节最大奖励与对手建模的自适应损失函数	采用经验回放池离线训练，学习的对手策略具有较大样本方差
	AMS-A3C[72]	在强化学习过程中，制订估计其他智能体策略的辅助任务	决策网络与模仿决策的对手模型共享结构、参数，降低模型学习成本	提出参数共享、策略表征2套方案，将对手建模融合进A3C算法	对手模型参数敏感，难以应对复杂场景、具有学习能力的对手
学习表征	PPO-Emb[74]	从交互样本中无监督地学习对手表征	提取同时具有策略提升效果和对手区分度的表征信息	无需领域知识，通用性强，适用大多数DRL算法	无法独立推断，用于辅助其他DRL算法决策
	RFM[73]	采用图网络学习智能体的社会关系表征	通过边缘属性、节点等图结构信息预测对手行动、评估对手社会关系强度	量化智能体交互的社会属性，网络结构具有较好拓展性	存在复杂交互关系的图网络计算困难
概率推理	P-BIT[76]	多智能体DRL的最优策略形式化为推理私有信息的概率下界	使用信念模块根据友方行为推理其私有信息	提出不完美信息条件下通过行动与队友传递私有信息的方法	适用于简单的二人合作场景
	ROMMEO[78]	多智能体DRL形式化为基于对手模型的最优策略变分推理	预测对手行动，用于实现学习最优策略的推理任务	提出最大熵目标的正则化的对手建模方法	在线优化参数，训练时间长。默认对手目标已知，无法适应未知智能体
自我–他人交互	SOM[79]	基于自身策略推理对手可能的目标，用于支撑决策	建立拟合对手策略的神经网络，通过优化对手策略反向推断对手的目标	无需额外模型和参数显式建模，仿照自身模型推理任意数量规模对手	智能体与对手共享目标，并且奖励结构取决于目标
	LOLA[81]	考虑具有学习能力的对手，解释对手学习参数的更新对自身策略影响	建模对手的价值函数，求其二阶导优化策略梯度	策略更新中增加了对手参数更新项，通过泰勒展开构造成高阶梯度项	默认对手使用可梯度优化的方法，并且无法察觉LOLA对其模型进行利用

4.1 游戏人工智能

游戏是对现实世界问题的抽象模拟，在游戏人工智能领域中，DRL-OM主要应用于为游戏智能体提供对手行动预测，提高智能体态势认知能力，进一步提高游戏局势判断的准确性，从而提升人机对抗水平。根据游戏种类的区别，对手建模可以分为棋牌类游戏、即时策略类游戏和街机格斗类游戏等不同应用类型。

棋牌类游戏具有不完美信息序贯决策的特性，因此这类游戏的智能体广泛采用基于反事实遗憾值最小化的显式建模方法。2017年1月加拿大阿尔伯特大学开发的DeepStack[82]在无限制德州扑克单挑赛中战胜所有参与的人类玩家。DeepStack采用了基于CFR理论的递归推理方法，通过强化学习值函数用于截断博弈子树，以对手遗憾值和自身范围解算残局。同年，卡耐基梅隆大学Brown

等[83]研发出的Libratus在无限制德州扑克单挑赛中击败了更强的人类职业选手。Libratus在游戏初始将博弈抽象表示为“蓝图”，以便基于蒙特卡罗CFR算法的求解。游戏过程的中后期，构建“蓝图”策略内计算的嵌套子博弈求解残局，并且针对可被对手利用的弱点提出自我改进。

在即时策略类游戏中，2019年10月研究机构DeepMind提出了星际争霸II人工智能AlphaStar[8]，面对异构智能体的策略学习问题，他们提出一种深度强化学习的联盟训练机制，从联盟中选取难以克服的对手作为陪练，发掘自身暴露的缺陷，以基于虚拟自博弈的显式DRL-OM方法学习历史最优策略。

对于街机格斗类游戏，2020年Tang等[84]提出自适应对手模型的滚动时域进化算法，采用交叉熵、Q学习和策略梯度等多种监督学习和强化学习等辅助任务的隐式建模方法优化预测对手行动的有效性。最终基于该算法的智能体击败了基于MCTS模型的智能体，在IEEE CoG举办的格斗游戏AI竞赛上取得最终冠军。此外，对手建模也广泛应用于第一人称射击[85-86]、足球机器人[87-89]等游戏的人工智能。

4.2 军事仿真

DRL-OM技术在军事仿真领域里具有广阔的发展前景。作为作战中辅助决策的手段，DRL-OM能够从战场获取的粗糙感知信息中精炼态势、意图，以增强判别意图遮蔽或战术欺骗的能力，从而能够为战场指挥控制节点提供高层决策支撑。广泛应用于水下航行器作战[90]、无人机集群对抗[91-93]、模拟蓝军对抗演习[94-96]等军事作战场景。

2016年，Dzieńkowski等[90]研究了考虑对手潜艇探测的自动水下机器人最优行动策略，提出一种应用于复杂海况的改进MiniMax均衡策略显式DRL-OM方法，利用神经网络模型灵活评估效用函数，并且由遗传算法调整模型参数。

2018年，空军工程大学的Huang等[91]在空战格斗中建立自主决策无人机飞行系统，利用贝叶斯推理和滚动时域优化(Bayesian inference and moving horizon optimization, BI&MHO)判断空战态势，并且基于机动惯性原理建立了对手飞行器的空中位置预测。2020年，空军工程大学的Zhou等[92]提出模拟人脑学习机制的智能空战学习框架，根据认知科学中对学习、知识和记忆的研究构建了大脑的认知机制模型。基于此模型和人类的推理能力，该方法建立了一个长期短期记忆的层次化多线程学习系统。该方法属于基于心智理论递归推理的隐式DRL-OM方法，相较BI&MHO具有更强的态势感知和对手动作预测能力。

2020年美国国防高级研究计划局(DARPA)提出“针对敌方战术的建设性机器学习作战”(constructive machine-learning battles with adversary tactics, COMBAT)项目[94]，旨在建立多组由机器学习生成战术的模拟蓝军分队，通过仿真作战环境充分挖掘士兵现有战术行为的缺陷并学习制定对策。同时，该项目结合深度强化学习算法和博弈论实现复杂战场环境决策，并利用OneSAF系统[95]评估人工智能系统在人机协同回路中的战术有效性。2021年，中科院的Liu等[96]将对手建模运用在兵棋推演中。该方法在静态地图添加注意力机制，赋予其动态特征，利用CNN网络提取战场态势表征，并使用位置预测模型根据长短时记忆网络预测兵力的行动轨迹，是一种典型的基于表征学习的隐式DRL-OM方法。

4.3 公共安全

在公共安全领域，DRL-OM的研究主要集中在博弈模型求解。公共安全可以看作是一类安防博弈问题，相较一般决策问题，安防博弈需要考虑智能化适应对手行为所带来的影响，表现出极强的对抗性和动态性。安防博弈问题与对手建模研究一脉相承[97]，其核心是通过博弈建模攻防双方的交互过程。根据公共安全的研究问题，对手建模的应用背景可以是交通网络安全[98-99]，区域防

恐威胁[100-101]、自动驾驶[102-103]、信息网络安全[104]等场景。

针对交通网络问题，Zhang等[98]提出了在道路网络中封锁犯罪分子逃逸路线的研究，设计一种封锁道路的攻击者–防御者安全博弈模型，并结合基于混合整数线性规划公式的最佳响应和基于高效双Oracle算法[58]的有效近似来计算最佳的防御策略。Xue等[99]将NFSP模型运用在大规模扩展式的网络安防博弈(network security games, NSGs)中，并使用斯坦伯格博弈的形式求解。首先，模型将NFSP中的最佳响应策略网络从行动–状态改造为行动–价值的映射，使NSGs中最佳响应的计算成为可能。随后，将NFSP智能体的平均策略网络转换为基于度量的分类器，为智能体提供高水平的行动，通过学习高效的图节点嵌入来利用NSGs图中包含的信息，模型在可扩展性和质量方面都有显著提升。

在区域防恐的应用中，Jain等[100]针对区域恐怖袭击研究城市警力资源分配，将问题建模成零和博弈，并使用双Oracle算法求解。现有研究缺乏动态调整资源分配的能力，因此，Zhang等[101]建立了一个模拟逃逸对手和有多个资源和实时信息的防御者之间交互过程的博弈模型，并设计增量策略生成算法求解。

此外，DRL-OM关注车辆自动驾驶的道路安全性问题。Tian等[102]研究了交叉路口交通场景，建立表示环形交叉口两车交汇的博弈论模型，提出了一种基于车辆交互模型并适应对手车辆驾驶员类型的环形交叉口自我决策算法。Notomista等[103]提出一种意识增强的纳什均衡搜寻方法解决车辆竞速的交互问题，采用避免碰撞的约束条件刻画与对手车辆的竞争关系，并且在对手模型存在不确定性下提出鲁棒性保护条件。

4.4 公共测试基准

针对深度强化学习领域内的通用问题，已有一系列学界公认的测试基准环境，如OpenAI GYM[105]、MuJoCo[106]、Fight ICE[107]等。基于这些开源框架，最先进的深度强化学习算法得以实现和对比。对于对手建模问题，特别是针对DRL-OM算法研究还没有形成统一的公共测试基准。

根据研究问题场景的具体要求，研究者提出了各式各样特性的实验环境。为了便于后续研究，本节根据场景特性对深度对手建模方法的实验场景进行分类：①按可观测信息的特性，可分为全局可观和部分可观实验场景；②按合作关系，实验场景可以分为合作(Cooperative)、竞争(Competitive)、混合(Mixed) 3种；③按智能体在环境中的行动顺序，实验场景可以分为同步决策、序贯决策实验场景。

跟据前文提及的可观测信息、合作关系、行动顺序等环境特征，从博弈论的角度，可将应用DRL-OM的问题分为扩展式博弈、马尔科夫博弈、团队马尔科夫博弈、部分可观马尔科夫过程、离散部分可观马尔科夫过程5种。

常见的实时环境仿真都是同步行动，即智能体间没有先后顺序，同时在环境中采取行动。此时若智能体能够完整观测状态，那么该博弈称为马尔科夫博弈(Markov game, MG)问题[108]，实验场景例如囚徒困境问题和硬币博弈等。在MG问题中，如果各个智能体通过合作来最大化共同的回报，将其称为团队马尔科夫博弈[109](team Markov game, Team MG)。

假如智能体的观测是局部的，对于利用个人目标实现集体回报最大化的，称为离散部分可观马尔科夫过程[110](decentralized partially observable Markov decision process, Dec-POMDP)，其他的博弈称为部分可观马尔科夫过程(POMDP)。棋牌类游戏常见的是序贯决策，当一方结束行动后，另一方开始行动，能够建模成扩展式博弈(extensive form games, EG)问题[111]。此外，按照状态、动作的表示也分为连续空间和离散空间。处理不同空间数量级的问题，适用的算法也相应不同。

在上述维度划分的基础上，表5汇总梳理了深度强化对手建模方法中常见的仿真场景、对应方法及其特性，并按照提出的博弈分类方法加以定义。为后续研究提供参考。

表5 常见实验场景、博弈模型、文献来源与问题特性

Table 5 Common experimental scenarios, game models, literature sources and problem characteristics

实验环境	博弈模型	文献	可观测信息	合作关系	行动顺序	状态动作
粒子世界	POMDP	[56-57,112-114]	部分可观	混合	同步	连续
德州扑克	EG	[38,40-42]	全局可观	竞争	序贯	离散
囚徒/硬币博弈	MG	[80]	全局可观	竞争	同步	离散
多智能体Mujoco	POMDP	[115]	部分可观	混合	同步	连续
网格世界	MG	[66-69]	全局可观	混合	同步	连续
迭代矩阵游戏	Team MG	[64,78]	全局可观	竞争	同步	离散
智力竞赛碗	EG	[26,71]	全局可观	竞争	序贯	离散
炸弹人	MG	[72]	全局可观	竞争	同步	离散
合作导航	Dec-POMDP	[63-64,116]	部分可观	合作	同步	离散
FightingICE	MG	[84,107]	全局可观	竞争	同步	连续
谷歌足球环境	POMDP	[87]	部分可观	混合	同步	连续

5 关键问题与发展方向

当前DRL-OM方法普遍面临着决策实时性差、难以应对非平稳性、模型学习效率低、无法有效地识别和使用欺骗以及对手利用等热点问题，为了在现实应用中更好地发挥作用，未来该领域可以从有模型强化学习、对手建模算法鲁棒性、对手快速适应性等方向开展研究。对此，本节针对关键问题和发展方向分别展开分析与探讨。

5.1 关键问题

5.1.1 实时性决策

现实应用中的即时决策场景十分常见，如竞技体育、战术决策等需要在短时间内做出反应决策。在强实时性的场景中，智能体面临着同时考虑计算准确性和效率的问题。Tang等[84]在滚动时域算法的基础上，使用深度强化学习预测对手行动，优化对手建模的有效性，实现了双人格斗游戏的博弈推理。然而，在多人在线游戏、大规模作战仿真等应用[10]中，面临的是多人实时博弈这种更为复杂的情况，实现这类场景下对手建模的实时性决策是亟需解决的难题。

5.1.2 非平稳性建模

非平稳性是指对手策略、环境状态转移、环境奖励不断发生变化，从而打破了MDP环境的马尔科夫性质，使得智能体无法学习到稳定的最优策略[117]。诸如部分可观测、环境动态变化、对手策略学习等因素都会导致智能体的学习环境非平稳。因此，非平稳性智能体需要权衡模型学习潜在误差所导致的不确定性和在不确定性下的对手模型利用。例如，智能体无法获取对手当前的状态和动作信息，只能利用自身局部信息进行推理。为此，研究者通过无监督学习[118-119]从自身观测中提取被建模者的历史轨迹特征，建立对手模型。Papoudakis等[112-113]基于局部信息建模，采用变分自编码器(variational autoencoders, VAEs)学习历史轨迹的策略表征，并将其用于A3C算法的决策过程。然而，在智能体共同学习、智能种群演化等复杂动态问题中[43]，策略难以收敛到静态最优，智能体陷入不断学习的循环。解决这类复杂任务中的非平稳性也是DRL-OM方法尚未解决的难题。

5.1.3 模型学习效率

如何快速准确地训练对手模型，提升数据利用率是对手建模的关键问题。一方面，DRL-OM

方法需要巨大的训练量和准确的奖励值作为指导，现实世界中的任务无法满足上述条件，因此，限制了应用场景[120-121]；另一方面，具有自适应能力的对手策略导致模型不断变化，交互数据具有时效性，快速适应对手动态变化是解决方法之一[122]。因此，如何避免学习过程消耗大量样本，从而以较少数据量在线学习到对手的策略是提升模型学习效率的关键问题。

5.1.4 欺骗与对手模型利用

智能体如何使用欺骗策略和如何识别对手欺骗策略是对手建模研究的关键问题。在棋牌游戏中，智能体采用诈唬策略诱使对手放弃应有的收益[123]。在即时策略类游戏中，通过“示弱”诱敌深入，放弃短期高收益的策略而采用伪装策略迷惑敌人[124]，这些都是典型的欺骗策略。Bontrager等[125]分析了基于DRL的智能体在欺骗性游戏中的表现，实验结果表明对比规划类方法，DRL方法无法有效地应对环境中的欺骗。尽管对手建模的子领域意图识别提供了诸多识别欺骗的解决方案[126-127]，但是当前的DRL-OM方法缺少对欺骗行为的研究，这一领域有很大的研究潜力。

对手利用(opponent exploitation)指发现对手有限理性的弱点，并根据对手偏离绝对理性的行动制定收益更高的策略。衡量对手策略可利用度的方式是对手策略与纳什均衡策略期望收益之间的差值[128]。Ganzfried等[129]提出博弈安全策略的表示形式，保证自身安全的同时利用对手的非最优策略，并在Kuhn扑克中超越学习纳什均衡策略的方法。DRL算法如何安全利用对手智能体，以及在更复杂场景计算对手利用策略，是DRL-OM领域值得探究的问题。

5.2 发展方向

5.2.1 有模型强化学习研究

在现有的对手建模方法中，自身智能体将其余智能体连同环境一起看作整体建模，忽视了对手模型在外部环境中的演化过程，无法有效地推理对手的行为。研究者们希望借助有模型强化学习显式地建立环境动力学模型[130-132]，将对手和环境之间的耦合进行拆解，帮助智能体更好地进行推理。

为此，Yu等[87]提出一种基于环境模型的贝叶斯递归推理方法，通过贝叶斯定理分配对手采取每种策略的权重，并将不同递归层次的策略组合构成对手策略，该算法在保证效率的同时避免了获取对手先验知识。Zhang等[132]将有模型强化学习的模型误差分为环境动态误差和对手模型误差，开发了自适应调整采样长度的对手建模误差控制方法，使用虚拟的对手和环境模型自演交互以扩充样本数据，达到高效利用数据的目的。

5.2.2 对手建模鲁棒性研究

DRL-OM方法通常需要智能体与对手进行长时间的交互训练，由于采样存在不稳定性，策略更新容易发生策略遗忘和崩塌的情况[133]。算法鲁棒性研究的是深度学习算法与环境进行交互时，如何保证策略不受限于局部最优以及策略不因采样误差而在更新中变差的问题[134-136]。除此之外，仿真环境与现实系统出现不匹配时，对手建模算法的鲁棒性决定了能否在模型迁移后保证其实用性[137-138]。

针对单个网络模型对策略表示能力不足的问题，Shen等[115]在学习对手策略时采用集成学习方法[139]训练多个网络共同表示当前对手模型，提高了模型的鲁棒性。除了集成学习思想，知识蒸馏也是提高模型准确性的一种方法[140]。知识蒸馏综合了多种专家网络，压缩网络结构信息，保留学习到的知识，经过知识蒸馏后的对手策略模型仍然保留了网络的对手特征提取能力[68,115]。

5.2.3 对手快速适应性研究

对手建模面临着难以适应具有学习能力的对手以及对新出现对手的重新适应等问题，元学习是实现智能体快速学习的途径[141-144]。一方面，它通过掌握已有的经验知识来指导新的任务，解决了重新训练和奖励函数构造的难题；另一方面，针对复杂动作空间的对手建模，元学习提供更高

层次的策略控制。例如，Kim 等[145]提出一种多智能体元策略梯度方法，将多智能体博弈过程建立为一种联合策略马尔科夫链，使用梯度优化主体智能体的初始参数，同时链式更新对其他智能体，达到快速适应对手的目的。

针对异构和不同机器学习方法的智能体，研究人员提出几种适应方法。Davies 等[114]提出学习模仿对手模型(learning to model opponent learning, LeMOL)，利用双向长短时记忆网络存储情节记忆，并模仿对手的学习过程。此外，Kim 等[145]考虑单个智能体的轨迹影响对手学习过程的情况，提供了可解释的异构智能体学习框架。

6 结论

本文旨在梳理深度强化学习算法在对手建模中的研究。首先，介绍了对手建模的研究背景，以及传统对手建模研究的不足之处，进一步介绍了基于深度强化学习的对手建模思想。随后，梳理了当前流行的深度强化学习算法，并且根据建模机理将 DRL-OM 方法划分为显式建模和隐式建模，分析了各类方法的思想、特性、局限和适用场景。针对深度强化学习的应用问题，本文汇总了游戏 AI、军事仿真、公共安全 3 个应用领域，梳理了现有研究的实验测试场景。最后，探讨了 DRL-OM 实现复杂现实应用面临的关键问题，并对 DRL-OM 方法的发展方向进行了展望。

参考文献：

[1] Rubinstein A. Modeling Bounded Rationality[M]. Cambridge, MA: MIT Press, 1998.

[2] Wang H, Kwong S, Jin Y, et al. Agent-based Evolutionary Approach for Interpretable Rule-based Knowledge Extraction[J]. IEEE Transactions on Systems Man and Cybernetics Part C(Applications and Reviews) (S1094-6977), 2005, 35(2): 143-155.

[3] Nguyen H V, Rezatofighi H, Vo B N, et al. Multi-Objective Multi-Agent Planning for Jointly Discovering and Tracking Mobile Objects[C]//AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020, 34 (5): 7227-7235.

[4] Sartoretti G, Kerr J, Shi Y, et al. PRIMAL: Pathfinding via Reinforcement and Imitation Multi-Agent Learning [J]. IEEE Robotics & Automation Letters (S2377-3766), 2019, 4(3): 2378-2385.

[5] Albrecht S V, Stone P. Autonomous Agents Modelling other Agents: A Comprehensive Survey and Open Problems[J]. Artificial Intelligence(S0004-3702), 2017, 258: 66-95.

[6] Volodymyr M, Koray K, David S, et al. Human-level Control through Deep Reinforcement Learning[J]. Nature (S0028-0836), 2019, 518(7540): 529-533.

[7] Enmin Z, Renye Y, Jinqiu L, et al. AlphaHoldem: High-Performance Artificial Intelligence for Heads-Up No-Limit Texas Hold'em from End-to-End Reinforcement Learning[C]//AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2022, 36(4): 4689-4697.

[8] 张蒙, 李凯, 吴哲, 等. 一种针对德州扑克AI的对手建模与策略集成框架[J]. 自动化学报, 2022, 48(4): 1004-1017.
Zhang Meng, Li Kai, Wu Zhe, et al. An Opponent Modeling and Strategy Integration Framework for Texas Hold'em[J]. Acta Automatica Sinica, 2022, 48(4): 1004-1017.

[9] Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster Level in StarCraft II Using Multi-agent Reinforcement Learning[J]. Nature(S0028-0836), 2019, 575(7782): 350-354.

[10] Bakkes S C J, Spronck P H M, Van Den Herik H J. Opponent Modelling for Case-Based Adaptive Game AI [J]. Entertainment Computing(S1875-9521), 2009, 1(1): 27-37.

[11] 罗俊仁, 张万鹏, 袁唯淋, 等. 面向多智能体博弈对抗的对手建模框架[J]. 系统仿真学报, 2022, 34(9): 1941-1955.
Luo Junren, Zhang Wanpeng, Yuan Weilin, et al. Research on Opponent Modeling Framework for Multi-agent Game Confrontation[J]. Journal of System Simulation, 2022, 34(9): 1941-1955.

[12] 刘婵娟, 赵天昊, 刘睿康, 等. 智能体对手建模研究进展[J]. 图学学报, 2021, 42(5): 703-711.
Liu Chanjuan, Zhao Tianhao, Liu Ruikang, et al. Research Progress of Opponent Modeling for Agent[J]. Journal of Graphics, 2021, 42(5): 703-711.

[13] Hernandez-Leal P, Kartal B, Taylor M E. A Survey and Critique of Multiagent Deep Reinforcement Learning [J]. Autonomous Agents and Multi-Agent Systems (S1387-2532), 2019, 33(6): 750-797.

[14] Tuyls K, Stone P. Multiagent Learning Paradigms[C]// Multi-Agent Systems and Agreement Technologies. 15th European Conference, EUMAS 2017, and 5th

International Conference, AT 2017. Evry, France: Springer International Publishing, 2018: 3-21.

[15] Nashed S, Zilberstein S. A Survey of Opponent Modeling in Adversarial Domains[J]. Journal of Artificial Intelligence Research(S1076-9757), 2022, 73: 277-327.

[16] Powers R, Shoham Y. Learning Against Opponents with Bounded Memory[C]//19th International Joint Conference on Artificial Intelligence. Edinburgh, Scotland: Morgan Kaufmann, 2005: 817-822.

[17] Chakraborty D, Stone P. Cooperating with A Markovian Ad Hoc Teammate[C]//2013 International Conference on Autonomous Agents and Multi-agent Systems. Beijing, China: AAMAS, 2013: 1085-1092.

[18] De Weerd H, Verbrugge R, Verheij B. Negotiating with other Minds: The Role of Recursive Theory of Mind in Negotiation with Incomplete Information[J]. Autonomous Agents and Multi-Agent Systems(S1387-2532), 2017, 31 (2): 250-287.

[19] Sonu E, Doshi P. Scalable Solutions of Interactive POMDPs Using Generalized and Bounded Policy Iteration[J]. Autonomous Agents and Multi-Agent Systems(S1387-2532), 2015, 29(3): 455-494.

[20] Zeng Y, Doshi P. Exploiting Model Equivalences for Solving Interactive Dynamic Influence Diagrams[J]. Journal of Artificial Intelligence Research(S1076-9757), 2012, 43: 211-255.

[21] Doshi P, Zeng Y, Chen Q. Graphical Models for Interactive POMDPs: Representations and Solutions[J]. Autonomous Agents and Multi-Agent Systems(S1387-2532), 2009, 18 (3): 376-416.

[22] Barrett S, Stone P. Cooperating with Unknown Teammates in Complex Domains: A Robot Soccer Case Study of ad Hoc Teamwork[C]//Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin, USA: AAAI 2015: 2010-2016.

[23] Erdogan C, Veloso M. Action Selection via Learning Behavior Patterns in Multi-Robot Systems[C]//Twenty-Second International Joint Conference on Artificial Intelligence. Barcelona, Spain: Morgan Kaufmann, 2011: 192-197.

[24] Weber B G, Mateas M. A Data Mining Approach to Strategy Prediction[C]//2009 IEEE Symposium on Computational Intelligence and Games. Milan, Italy: IEEE, 2009: 140-147.

[25] Schadd F, Bakkes S, Spronck P. Opponent Modeling in Real-Time Strategy Games[C]//The 8th Annual European Game-On Conference on Simulation and AI in Computer Games(GAMEON). Bologna, Italy: Marco Roccetti, 2007: 61-70.

[26] He H, Boyd-Graber J, Kwok K, et al. Opponent Modeling in Deep Reinforcement Learning[C]// International Conference on Machine Learning. New York, USA: PMLR, 2016: 1804-1813.

[27] Baker C, Saxe R, Tenenbaum J. Bayesian Theory of Mind: Modeling Joint Belief-desire Attribution[C]//Annual Meeting of the Cognitive Science Society. Boston, USA: Cognitive Science Society, 2011: 2469-2474.

[28] Fern A, Tadepalli P. A Computational Decision Theory for Interactive Assistants[C]//23rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc, 2010: 577-585.

[29] Sohrabi S, Riabov A V, Udrea O. Plan Recognition as Planning Revisited[C]//Twenty-Fifth International Joint Conference on Artificial Intelligence. New York, USA: AAAI, 2016: 3258-3264.

[30] Albrecht S V, Stone P. Reasoning about Hypothetical Agent Behaviours and Their Parameters[C]//International Conference on Autonomous Agents and Multiagent Systems. Richland, USA: International Foundation for Autonomous Agents and Multiagent Systems, 2017: 547-555.

[31] Albrecht S V, Crandall J W, Ramamoorthy S. Belief and Truth in Hypothesised Behaviours[J]. Artificial Intelligence (S0004-3702), 2016, 235: 63-94.

[32] 孙长银, 穆朝絮. 多智能体深度强化学习的若干关键科学问题[J]. 自动化学报, 2020, 46(7): 1301-1312. DOI: 10.16383/j.aas.c200159.
Sun Changyin, Mu Chaoxu. Important Scientific Problems of Multi-Agent Deep Reinforcement Learning [J]. Acta Automatica Sinica, 2020, 46(7): 1301-1312. DOI:10.16383/j.aas.c200159.

[33] Hausknecht M, Stone P. Deep Recurrent Q-Learning for Partially Observable MDPs[C]//2015 AAAI Fall Symposium Series. Arlington, Virginia: AAAI, 2015.

[34] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous Control with Deep Reinforcement Learning[J/OL]. ArXiv Preprint ArXiv: 1509.02971, 2015. [2022-03-25]. https://arxiv.org/abs/1509.02971.

[35] Mnih V. Asynchronous Methods for Deep Reinforcement Learning[C]//33rd International Conference on International Conference on Machine Learning. New York, USA: PMLR, 2016: 1928-1937.

[36] Schulman J, Levine S, Moritz P, et al. Trust Region Policy Optimization[C]//32nd International Conference on International Conference on Machine Learning. Lille, France: PMLR, 2015: 1889-1897.

[37] Schulman J, Wolski F, Dhariwal P, et al. Proximal Policy

Optimization Algorithms[J/OL]. ArXiv Preprint ArXiv: 1707.06347, 2017. [2022-03-25]. https://arxiv. org/abs/1707.06347.

[38] Heinrich J, Lanctot M, Silver D. Fictitious Self-play in Extensive-form Games[C]//32nd International Conference on International Conference on Machine Learning. Lille, France: PMLR, 2015: 805-813.

[39] Heinrich J, Silver D. Deep Reinforcement Learning from Self-play in Imperfect-Information Games[J/OL]. ArXiv Preprint ArXiv: 1603.01121, 2016. [2022-03-25]. https://arxiv.org/abs/1603.01121.

[40] Kawamura K, Tsuruoka Y. Neural Fictitious Self-play on ELF Mini-rts[J/OL]. ArXiv Preprint ArXiv: 1902.02004, 2019. [2022-03-25]. https://arxiv.org/abs/1902.02004.

[41] Zhang L, Chen Y, Wang W, et al. A Monte Carlo Neural Fictitious Self-Play Approach to Approximate Nash Equilibrium in Imperfect-information Dynamic Games [J]. Frontiers of Computer Science(S2095-2228), 2021, 15(5): 1-14.

[42] Lanctot M, Zambaldi V, Gruslys A, et al. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning[C]//31st Conference on Neural Information Processing Systems(NIPS). Long Beach, USA: MIT Press, 2017: 4193-4206.

[43] Omidshafiei S, Papadimitriou C, Piliouras G, et al. α-rank: Multi-agent Evaluation by Evolution[J]. Scientific Reports(S2045-2322), 2019, 9(1): 1-29.

[44] Muller P, Omidshafiei S, Rowland M, et al. A Generalized Training Approach for Multiagent Learning[J/OL]. ArXiv Preprint ArXiv: 1909.12823, 2019. [2022-03-25]. https://arxiv.org/abs/1909.12823.

[45] Balduzzi D, Garnelo M, Bachrach Y, et al. Open-ended Learning in Symmetric Zero-sum Games[C]// International Conference on Machine Learning. Long Beach, USA: PMLR, 2019: 434-443.

[46] Martin Z, Michael J, Michael B, et al. Regret Minimization in Games with Incomplete Information[C]// 20th International Conference on Neural Information Processing Systems(NIPS'07). Red Hook, New York, USA: Curran Associates Inc, 2007: 1729-1736.

[47] Gibson R. Regret Minimization in Games and the Development of Champion Multiplayer Computer Poker-Playing Agents[D]. Edmonton: University of Alberta, 2014.

[48] Johanson M, Bard N, Lanctot M, et al. Efficient Nash Equilibrium Approximation through Monte Carlo Counterfactual Regret Minimization[C]//International Conference on Autonomous Agents and Multiagent Systems. Valencia, Spain: International Foundation for Autonomous Agents and Multiagent Systems, 2012: 837-846.

[49] Lanctot M. Monte Carlo Sampling and Regret Minimization for Equilibrium Computation and Decision-Making in Large Extensive Form Games[D]. Edmonton: University of Alberta, 2013.

[50] Brown N, Sandholm T. Reduced Space and Faster Convergence in Imperfect-information Games via Pruning[C]//International Conference on Machine Learning. Sydney, Australia: PMLR, 2017: 596-604.

[51] Tammelin O. Solving Large Imperfect Information Games Using CFR+[J/OL]. ArXiv Preprint ArXiv: 1407.5042, 2014. [2022-03-25]. https://arxiv.org/abs/1407.5042.

[52] Gilpin A, Sandholm T. Better Automated Abstraction Techniques for Imperfect Information Games, with Application to Texas Hold'em Poker[C]//International Joint Conference on Autonomous Agents and Multiagent Systems. New York, USA: International Foundation for Autonomous Agents and Multiagent Systems, 2007: 1-8.

[53] Waugh K, Schnizlein D, Bowling M, et al. Abstraction Pathologies in Extensive Games[C]//8th International Conference on Autonomous Agents and Multiagent Systems. Budapest, Hungary: International Foundation for Autonomous Agents and Multiagent Systems, 2009: 781-788.

[54] 王鹏程. 基于深度强化学习的非完备信息机器博弈研究[D]. 哈尔滨: 哈尔滨工业大学, 2016.
Wang Pengcheng. Research on Imperfect Information Machine Game Based on Deep Reinforcement Learning [D]. Harbin: Harbin Institute of Technology, 2016.

[55] Wright J R, Leyton-Brown K. Level-0 Models for Predicting Human Behavior in Games[J]. Journal of Artificial Intelligence Research(S1076-9757), 2019, 64: 357-383.

[56] Li S, Wu Y, Cui X, et al. Robust Multi-agent Reinforcement Learning via Minimax Deep Deterministic Policy Gradient[C]//AAAI Conference on Artificial Intelligence. Hawaii, USA: AAAI, 2019, 33: 4213-4220.

[57] Lowe R, Wu Y, Tamar A, et al. Multi-agent Actor-critic for Mixed Cooperative-competitive Environments[C]// 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc, 2017: 6382-6393.

[58] McMahan H B, Gordon G J, Blum A. Planning in the Presence of Cost Functions Controlled by an Adversary [C]//20th International Conference on Machine Learning (ICML-03). Helsinki, Finland: ACM, 2003: 536-543.

[59] Frith C, Frith U. Theory of Mind[J]. Current Biology

(S0960-9822), 2001, 15(17): R644-R645.

[60] Rabinowitz N C, Perbet F, Song H F, et al. Machine Theory of Mind[C]//35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018: 4218-4227.

[61] Aucher G, Bolander T. Undecidability in Epistemic Planning[C]//23rd International Joint Conference on Artificial Intelligence. Beijing, China: Morgan Kaufmann, 2013: 27-33.

[62] Hartford J, Wright J R, Leyton-Brown K. Deep Learning for Predicting Human Strategic Behavior[C]//30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc, 2016: 2432-2440.

[63] Wen Y, Yang Y, Luo R, et al. Probabilistic Recursive Reasoning for Multi-agent Reinforcement Learning[J/OL]. ArXiv Preprint ArXiv: 1901.09207, 2019. [2022-03-25]. https://arxiv.org/abs/1901.09207.

[64] Wen Ying, Yang Yaodong, Luo Rui, et al. Modelling Bounded Rationality in Multi-agent Interactions by Generalized Recursive Reasoning[C]//Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20). Yokohama, Japan: AAAI, 2020: 414-421.

[65] Annie W, Thomas B, Anna V, et al. Multiagent Deep Reinforcement Learning: Challenges and Directions Towards Human-Like Approaches[J]. Artificial Intelligence Review(S0269-2821), 2022. DOI: 10.1007/s10462-022-10299-x. [2022-03-25]. https://link. springer.com/article/10.1007/s10462-022-10299-x.

[66] Everett R, Roberts S. Learning Against Non-stationary Agents with Opponent Modelling and Deep Reinforcement Learning[C]//2018 AAAI Spring Symposium Series. Palo Alto, California: AAAI, 2018.

[67] Hernandez-Leal P, Rosman B, Taylor M E, et al. A Bayesian Approach for Learning and Tracking Switching, Non-stationary opponents[C]//2016 International Conference on Autonomous Agents & Multiagent Systems. Singapore, Singapore: International Foundation for Autonomous Agents and Multiagent Systems, 2016: 1315-1316.

[68] Zheng Y, Meng Z, Hao J, et al. A Deep Bayesian Policy Reuse Approach against Non-stationary Agents[C]//32nd International Conference on Neural Information Processing Systems. Red Hook, New York: Curran Associates Inc, 2018: 962-972.

[69] Yang T, Hao J, Meng Z, et al. Towards Efficient Detection and Optimal Response Against Sophisticated Opponents[C]//28th International Joint Conference on Artificial Intelligence. Macao, China: AAAI, 2019: 623-629.

[70] Lample G, Chaplot D S. Playing FPS Games with Deep Reinforcement Learning[C]//Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, California, USA: AAAI, 2017: 2140-2146.

[71] Hong Z W, Su S Y, Shann T Y, et al. A Deep Policy Inference Q-Network for Multi-agent Systems[C]//17th International Conference on Autonomous Agents and Multi-Agent Systems. Stockholm, Sweden: International Foundation for Autonomous Agents and Multiagent Systems, 2018: 1388-1396.

[72] Hernandez-Leal P, Kartal B, Taylor M E. Agent Modeling as Auxiliary Task for Deep Reinforcement Learning[C]//AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. Palo Alto, California: AAAI, 2019, 15(1): 31-37.

[73] Tacchetti A, Song H F, Mediano P A M, et al. Relational Forward Models for Multi-agent Learning[J/OL]. ArXiv Preprint ArXiv: 1809.11044, 2018. [2022-03-25]. https://arxiv.org/abs/1809.11044.

[74] Grover A, Al-Shedivat M, Gupta J, et al. Learning Policy Representations in Multiagent Systems[C]//37th International Conference on Machine Learning Conference. Vienna, Austria: PMLR, 2018: 1802-1811.

[75] Ha D, Schmidhuber J. Recurrent World Models Facilitate Policy Evolution[C]//32nd International Conference on Neural Information Processing Systems. Red Hook, New York, USA: Curran Associates Inc, 2018: 2455-2467.

[76] Tian Z, Zou S, Davies I, et al. Learning to Communicate Implicitly by Actions[C]//AAAI Conference on Artificial Intelligence. New York: AAAI, 2020, 34(5): 7261-7268.

[77] Haarnoja T, Zhou A, Hartikainen K, et al. Soft Actor-critic Algorithms and Applications[J/OL]. ArXiv Preprint ArXiv:1812.05905, 2018. [2022-03-25]. https://arxiv.org/abs/1812.05905.

[78] Zheng T, Ying W, Zhichen G, et al. A Regularized Opponent Model with Maximum Entropy Objective [J/OL]. ArXiv Preprint ArXiv: 1905.08087, 2019. [2022-03-25]. https://arxiv.org/abs/1905.08087.

[79] Raileanu R, Denton E, Szlam A, et al. Modeling Others Using Oneself in Multi-agent Reinforcement Learning [C]//International Conference on Machine Learning. Vienna, Austria: PMLR, 2018: 4257-4266.

[80] Zhang C, Lesser V. Multi-agent Learning with Policy Prediction[C]//Twenty-Fourth AAAI Conference on Artificial Intelligence. Atlanta, USA: AAAI, 2010: 927-934.

[81] Foerster J N, Chen R Y, Al-Shedivat M, et al. Learning with Opponent-learning Awareness [C]//17th

International Conference on Autonomous Agents and MultiAgent Systems. Stockholm, Sweden: International Foundation for Autonomous Agents and Multiagent Systems, 2018: 122-130.

[82] Matej Moravík, Schmid M, Burch N, et al. DeepStack: Expert-Level Artificial Intelligence in No-Limit Poker[J]. Science(S0036-8075), 2017, 356(6337): 508-513.

[83] Brown N, Sandholm T. Superhuman AI for Heads-up No-limit Poker: Libratus Beats Top Professionals[J]. Science (S0036-8075), 2017, 359(6374): 418-424.

[84] Tang Z, Zhu Y, Zhao D. Enhanced Rolling Horizon Evolution Algorithm with Opponent Model Learning[J]. IEEE Transactions on Games(S2475-1502), 2020. DOI: 10.1109/TG. 2020.3022698. [2022-03-25]. https://ieeexplore.ieee.org/document/9190073.

[85] Huang S, Su H, Zhu J, et al. Combo-action: Training Agent for FPS Game with Auxiliary Tasks[C]//AAAI Conference on Artificial Intelligence. Hawaii, USA: AAAI, 2019, 33(1): 954-961.

[86] Wang K, Chang K C, Chang Z W. Determinants of We-intention for Continue Playing FPS Game: Cooperation and Competition[C]//7th Multidisciplinary in International Social Networks Conference and the 3rd International Conference on Economics, Management and Technology. Kaohsiung, Taiwan, China: ACM, 2020: 1-9.

[87] Yu X, Jiang J, Jiang H, et al. Model-based Opponent Modeling[J/OL]. ArXiv Preprint ArXiv: 2108.01843, 2021. [2022-03-25]. https://arxiv.org/abs/2108.01843.

[88] Iglesias J A, Ledezma A, Sanchis A. Opponent Modeling in RoboCup Soccer Simulation[C]//Workshop of Physical Agents. Cham: Springer, 2018: 303-316.

[89] Wu Z, Li K, Zhao E, et al. L2e: Learning to Exploit Your Opponent[J/OL]. ArXiv Preprint ArXiv: 2102.09381, 2021. [2022-03-25]. https://arxiv.org/abs/2102.09381.

[90] Dzieńkowski B J, Strode C, Markowska-Kaczmar U. Employing Game Theory and Computational Intelligence to Find the Optimal Strategy of an Autonomous Underwater Vehicle Against A Submarine[C]//2016 Federated Conference on Computer Science and Information Systems(FedCSIS). Gdansk, Poland: IEEE, 2016: 31-40.

[91] Changqiang H, Kangsheng D, Hanqiao H, et al. Autonomous Air Combat Maneuver Decision Using Bayesian Inference and Moving Horizon Optimization [J]. Journal of Systems Engineering and Electronics (S1004-4132), 2018, 29(1): 86-97.

[92] Zhou K, Wei R, Zhang Q, et al. Learning System for Air Combat Decision Inspired by Cognitive Mechanisms of the Brain[J]. IEEE Access(S2169-3536), 2020, 8: 8129-8144.

[93] 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. DOI:10.16383/j.aas.c201059.
Shi Wei, Feng Yanghe, Cheng Guangquan, et al. Research on Multi-aircraft Cooperative Air Combat Method Based on Deep Reinforcement Learning[J]. Acta Automatica Sinica, 2021, 47(7): 1610-1623. DOI: 10.16383/j.aas.c201059.

[94] Defense Advanced Research Projects Agency. Constructive Machine-learning Battles with Adversary-Tactics[EB/OL]. (2021-07-21)[2022-03-28]. https://www.darpa.mil/program/constructive-machine-learning-battles-with-adversary-tactics.

[95] Parsons D, Surdu J, Jordan B. OneSAF: a next Generation Simulation Modeling the Contemporary Operating Environment[C]//Euro-simulation Interoperability Workshop. Toulouse, France: Simulation Interoperability Standards Organization (SISO), 2005: 27-29.

[96] Liu X, Zhao M, Dai S, et al. Tactical Intention Recognition in Wargame[C]//2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS). Chengdu, China: IEEE, 2021: 429-434.

[97] 王震, 袁勇, 安波, 等. 安全博弈论研究综述[J]. 指挥与控制学报, 2015, 1(2): 121-149.
Wang Zhen, Yuan Yong, An Bo, et al. An Overview of Security Games[J]. Journal of Command and Control, 2015, 1(2): 121-149.

[98] Zhang Y, An B, Tran-Thanh L, et al. Optimal Escape Interdiction on Transportation Networks[C]//26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: AAAI, 2017: 3936-3944.

[99] Xue W, Zhang Y, Li S, et al. Solving Large-scale Extensive-form Network Security Games via Neural Fictitious Self-play[J/OL]. ArXiv Preprint ArXiv: 2106.00897, 2021. [2022-03-25]. https://arxiv. org/abs/2106.00897.

[100] Jain M, Korzhyk D, Vaněk O, et al. A Double Oracle Algorithm for Zero-sum Security Games on Graphs[C]// The 10th International Conference on Autonomous Agents and Multiagent Systems. Taipei, Taiwan, China: International Foundation for Autonomous Agents and Multiagent Systems, 2011: 327-334.

[101] Zhang Y, Guo Q, An B, et al. Optimal Interdiction of Urban Criminals with the Aid of Real-time Information [C]//AAAI Conference on Artificial Intelligence. Hawaii, USA: AAAI, 2019, 33(1): 1262-1269.

[102] Tian R, Li S, Li N, et al. Adaptive Game-theoretic Decision Making for Autonomous Vehicle Control at

Roundabouts[C]//2018 IEEE Conference on Decision and Control(CDC). Miami Beach, USA: IEEE, 2018: 321-326.

[103] Notomista G, Wang M, Schwager M, et al. Enhancing Game-theoretic Autonomous Car Racing Using Control Barrier Functions[C]//2020 IEEE International Conference on Robotics and Automation(ICRA). Paris, France: IEEE, 2020: 5393-5399.

[104] Okamoto S, Hazon N, Sycara K. Solving Non-zero Sum Multiagent Network Flow Security Games with Attack Costs[C]//11th International Conference on Autonomous Agents and Multiagent Systems. Valencia, Spain: International Foundation for Autonomous Agents and Multiagent Systems, 2012(2): 879-888.

[105] Brockman G, Cheung V, Pettersson L, et al. Openai Gym[J/OL]. ArXiv Preprint ArXiv: 1606.01540, 2016. [2022-03-25]. https://arxiv.org/abs/1606.01540.

[106] Todorov E, Erez T, Tassa Y. Mujoco: A Physics Engine for Model-based Control[C]//2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vilamoura, Portugal: IEEE, 2012: 5026-5033.

[107] Lu F, Yamamoto K, Nomura L H, et al. Fighting Game Artificial Intelligence Competition Platform[C]//2013 IEEE 2nd Global Conference on Consumer Electronics (GCCE). Tokyo, Japan: IEEE, 2013: 320-323.

[108] Littman M L. Markov Games as A Framework for Multi-agent Reinforcement Learning[M]// Machine Learning Proceedings. San Mateo, CA: Morgan Kaufmann, 1994: 157-163.

[109] Wang X, Sandholm T. Reinforcement Learning to Play An Optimal Nash Equilibrium in Team Markov Games [C]//15th International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2002: 1603-1610.

[110] Monahan G E. State of the Art-A Survey of Partially Observable Markov Decision Processes: Theory, Models, and Algorithms[J]. Management Science(S0025-1909), 1982, 28(1): 1-16.

[111] Kuhn H W, Tucker A W. Contributions to the Theory of Games[M]. Princeton, New Jersey: Princeton University Press, 1953.

[112] Papoudakis G, Albrecht S V. Variational Autoencoders for Opponent Modeling in Multi-agent Systems[J/OL]. ArXiv Preprint ArXiv: 2001.10829, 2020. [2022-03-25]. https://arxiv.org/abs/2001.10829.

[113] Papoudakis G, Christianos F, Albrecht S. Agent Modelling under Partial Observability for Deep Reinforcement Learning[C]//Advances in Neural Information Processing Systems. Virtual: Curran Associates Inc. 2021, 34: 19210-19222.

[114] Davies I, Tian Z, Wang J. Learning to Model Opponent Learning[C]//AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020, 34(10): 13771-13772.

[115] Shen M, How J P. Robust Opponent Modeling via Adversarial Ensemble Reinforcement Learning[C]// International Conference on Automated Planning and Scheduling. Guangzhou, China: AAAI, 2021, 31: 578-587.

[116] Wang T, Bao X, Clavera I, et al. Benchmarking Model-based Reinforcement Learning[J/OL]. ArXiv Preprint ArXiv: 1907.02057, 2019. [2022-03-25]. https://arxiv.org/abs/1907.02057.

[117] Papoudakis G, Christianos F, Rahman A, et al. Dealing with Non-stationarity in Multi-agent Deep Reinforcement Learning[J/OL]. ArXiv Preprint ArXiv: 1906.04737, 2019. [2022-03-25]. https://arxiv. org/abs/1906.04737.

[118] Cemgil T, Ghaisas S, Dvijotham K, et al. The Autoencoding Variational Autoencoder[C]//Advances in Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc. 2020, 33: 15077-15087.

[119] Hastie T, Friedman J, Tibshirani R. The Elements of Statistical Learning: Data mining, Inference, and Prediction[M]. Berlin, German: Springer, 2001.

[120] Genc S, Mallya S, Bodapati S, et al. Zero-shot Reinforcement Learning with Deep Attention Convolutional Neural Networks[J/OL]. ArXiv Preprint ArXiv: 2001.00605, 2020. [2022-03-25]. https://arxiv.org/abs/2001.00605.

[121] Wang Y, Yao Q, Kwok J T, et al. Generalizing from a Few Eamples: A Survey on Few-shot Learning[J]. ACM Computing Surveys(CSUR) (S0360-0300), 2020, 53(3): 1-34.

[122] Finn C, Abbeel P, Levine S. Model-agnostic Meta-Learning for Fast Adaptation of Deep Networks[C]// International Conference on Machine Learning. Sydney, Australia: PMLR, 2017: 1126-1135.

[123] Southey F, Bowling M, Larson B, et al. Bayes' Bluff: Opponent Modelling in Poker[C]//Twenty-first Conference on Uncertainty in Artificial Intelligence. Edinburgh, Scotland: AUAI, 2005: 550-558.

[124] 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/OL]. ArXiv preprint ArXiv: 1703.10069, 2017. [2022-03-25]. https://arxiv.org/abs/1703.10069.

[125] Bontrager P, Khalifa A, Anderson D, et al.

"Superstition" in the Network: Deep Reinforcement Learning Plays Deceptive Games[C]//AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. Atlanta, USA: AAAI, 2019, 15(1): 10-16.

[126] Johnson P E, Grazioli S, Jamal K. Fraud detection: Intentionality and Deception in Cognition[J]. Accounting, Organizations and Society(S0361-3682), 1993, 18(5): 467-488.

[127] Masters P, Kirley M, Smith W. Extended Goal Recognition: A Planning-based Model for Strategic Deception[C]//20th International Conference on Autonomous Agents and Multi-agent Systems. Virtual Event, United Kingdom: International Foundation for Autonomous Agents and Multiagent Systems, 2021: 871-879.

[128] Wang Z, Boularias A, Mülling K, et al. Balancing Safety and Exploitability in Opponent Modeling[C]//AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI, 2011, 25(1): 1515-1520.

[129] Ganzfried S, Sandholm T. Safe Opponent Exploitation [J]. ACM Transactions on Economics and Computation (TEAC)(S2167-8375), 2015, 3(2): 1-28.

[130] Deisenroth M, Rasmussen C E. PILCO: A Model-based and Data-efficient Approach to Policy Search[C]//28th International Conference on Machine Learning (ICML-11). Kyoto, Japan: PMLR, 2011: 465-472.

[131] Chua K, Calandra R, Mc Allister R, et al. Deep Reinforcement Learning in a Handful of Trials Using Probabilistic Dynamics Models[C]//32nd International Conference on Neural Information Processing Systems. Red Hook, New York, USA: Curran Associates Inc, 2018: 4759-4770.

[132] Zhang W, Wang X, Shen J, et al. Model-based Multi-agent Policy Optimization with Adaptive Opponent-wise Rollouts[J/OL]. ArXiv preprint ArXiv: 2105.03363, 2021. [2022-03-25]. https://arxiv.org/abs/2105.03363.

[133] Pan X, Seita D, Gao Y, et al. Risk Averse Robust Adversarial Reinforcement Learning[C]//2019 International Conference on Robotics and Automation (ICRA). Montreal, Canada: IEEE, 2019: 8522-8528.

[134] Vinitsky E, Du Y, Parvate K, et al. Robust Reinforcement Learning Using Adversarial Populations [J/OL]. ArXiv Preprint ArXiv: 2008.01825, 2020. [2022-03-25]. https://arxiv.org/abs/2008.01825.

[135] Ramoni M, Sebastiani P. Robust Learning with Missing Data[J]. Machine Learning(S0885-6125), 2001, 45(2): 147-170.

[136] Steinhardt J. Robust Learning: Information Theory and Algorithms[M]. Palo Alto, CA: Stanford University, 2018.

[137] Pinto L, Davidson J, Sukthankar R, et al. Robust Adversarial Reinforcement Learning[C]//International Conference on Machine Learning. Sydney, Australia: PMLR, 2017: 2817-2826.

[138] Shioya H, Iwasawa Y, Matsuo Y. Extending Robust Adversarial Reinforcement Learning Considering Adaptation and Diversity[J/OL]. ArXiv Preprint ArXiv: 1703.02702, 2017. [2022-03-25]. https://arxiv. org/abs/1703.02702.

[139] Sagi O, Rokach L. Ensemble Learning: A Survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery(S1942-4787), 2018, 8(4): e1249.

[140] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network[J/OL]. ArXiv Preprint ArXiv: 1503.02531, 2015. [2022-03-25]. https://arxiv. org/abs/1503.02531.

[141] Wang J X, Kurth-Nelson Z, Kumaran D, et al. Prefrontal Cortex as a Meta-reinforcement Learning System[J]. Nature Neuroscience(S1097-6256), 2018, 21(6): 860-868.

[142] Nagabandi A, Clavera I, Liu S, et al. Learning to Adapt in Dynamic, Real-world Environments through Meta-reinforcement Learning[J/OL]. ArXiv Preprint ArXiv: 1803.11347, 2018. [2022-03-25]. https://arxiv. org/abs/1803.11347.

[143] Yu T, Quillen D, He Z, et al. Meta-world: A Benchmark and Evaluation for Multi-task and Meta Reinforcement Learning[C]//Conference on Robot Learning. Virtual Event/Cambridge, USA: PMLR, 2020: 1094-1100.

[144] Gupta A, Mendonca R, Liu Y X, et al. Meta-reinforcement Learning of Structured Exploration Strategies[C]//32nd International Conference on Neural Information Processing Systems. Montréal Canada: Curran Associates Inc, 2018: 5307-5316.

[145] Kim D K, Liu M, Riemer M D, et al. A Policy Gradient Algorithm for Learning to Learn in Multiagent Reinforcement Learning[C]//International Conference on Machine Learning. Virtual Event: PMLR, 2021: 5541-5550.