生成式人工智能中的强化学习鲁棒性研究进展

刘自轩[1]，赵乙[2]，徐恪[1,3*]

（1. 清华大学计算机科学与技术系，北京 100084；
2. 北京理工大学网络空间安全学院，北京 100081；
3. 中关村实验室，北京 100194）

摘要：随着生成式人工智能的蓬勃发展，强化学习作为一项关键技术，已被广泛应用于提升模型的综合性能。然而，强化学习的引入也将模型的决策核心暴露在新的攻击面之下，其涉及的状态感知、奖励评估和动作执行每一环节均可能存在安全风险，使得强化学习过程的鲁棒性成为至关重要的安全问题。因此，从强化学习鲁棒性的视角对该交叉领域进行了系统性综述，通过建立传统强化学习安全与新兴应用场景之间的联系，深入剖析了各类攻击算法向对抗性提示、奖励模型投毒等新型威胁的演化路径；与此对应，从对抗训练、博弈论、可验证鲁棒性等多个维度，探讨了经典防御策略的迁移潜力。最后，对未来研究方向作出展望，旨在为构建更安全、可靠的生成式人工智能系统提供理论参考与技术洞见。

关键词：生成式人工智能；深度强化学习；深度学习安全；对抗攻击；算法鲁棒性

中图分类号：TP391 **文献标志码：**A

DOI: 10.20172/j.issn.2097-3136.250604

Advances in robustness of reinforcement learning in generative AI

Liu Zixuan[1], Zhao Yi[2], Xu Ke[1,3*]

（1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084；
2. School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081；
3. Zhongguancun Lab, Beijing 100194）

Abstract: As generative AI continues its rapid development, Reinforcement Learning (RL) has emerged as a key technology for improving the overall performance of models. However, the introduction of RL has also exposed the decision-making core of models to new attack surfaces, since every stage including state perception, reward evaluation, and action execution may entail security risks. Such a phenomenon renders the robustness of the RL process a vital security issue. A systematic review of this interdisciplinary domain, i.e., the intersection of generative AI and RL, is presented from the perspective of RL robustness. By establishing connections between traditional RL security and emerging application scenarios, the evolution of classic attacks against the core elements of RL into novel threats such as adversarial prompts and reward model poisoning is analyzed. Correspondingly, the migration potential of classic defense strategies is explored from multiple dimensions, namely adversarial training, game theory, and verifiable robustness. Finally, future research directions are discussed, with the aim of providing theoretical references and technical insights for building more secure and reliable generative AI systems.

Keywords: generative AI; deep reinforcement learning; deep learning security; adversarial attack; algorithm robustness

*** 通信作者：**E-mail: xuke@tsinghua.edu.cn

基金项目：国家杰出青年科学基金（62425201）；国家自然科学基金（62132011）

引用格式：刘自轩，赵乙，徐恪. 生成式人工智能中的强化学习鲁棒性研究进展 [J]. 网络空间安全科学学报，2025，3（6）：55 - 67.

Citation Format: Liu Z X，Zhao Y，Xu K. Advances in robustness of reinforcement learning in generative AI[J]. Journal of Cybersecurity，2025，3（6）：55 - 67.

0 引言

近年来，以大语言模型（Large Language Model，LLM）为代表的生成式人工智能技术实现了革命性的突破，并在自然语言处理、代码生成、人机交互等领域展现出显著的潜力与活力。然而，当这类强大的模型被广泛集成到社会生产与日常生活中时，其安全性和鲁棒性问题也日益凸显，成为技术发展的关键短板。研究表明，攻击者可通过精心设计的输入，即“提示”，来利用大语言模型的内在缺陷。例如，越狱攻击[1]通过欺骗性或逻辑陷阱类提示，诱使模型生成有害、违规或虚假的内容；而提示注入攻击[2]则通过在输入中植入恶意指令的方式，劫持模型的原始意图，使其执行攻击者设定的任务。这些攻击手段凸显了生成式 AI 在鲁棒性方面的脆弱性，表明确保模型在复杂对抗环境中行为的可靠性和可控性，是一项极其重要且充满挑战的任务。

为提升模型的性能与安全性，强化学习（Reinforcement Learning, RL）正成为关键赋能技术。在大型语言模型领域，人类反馈强化学习（Reinforcement Learning from Human Feedback，RLHF）通过奖励模型和策略优化，已成为实现模型与人类价值观的对齐的主流方法[3-4]。在文生图模型（Text-to-Image Models）中，强化学习被用于根据美学评分或用户偏好来优化图像的生成质量[5]。在 AI 代理应用中，强化学习则帮助模型完成复杂且多步骤的决策任务[6]。然而，对强化学习的深度依赖程度，也意味着将模型暴露在全新的攻击面之下。无论是奖励信号的设计、训练数据的收集，还是策略优化的过程，都可能成为攻击者可利用的薄弱点，直接威胁生成模型的决策核心。

事实上，生成式 AI 所面临的强化学习安全威胁，在很大程度上可视为传统强化学习安全问题在新场景下的具体体现与演化。例如，在传统强化学习中常见的基于状态的攻击，在应用了强化学习的语言模型中，已演变为通过对抗性提示进行越狱的攻击方式。在这种场景下，攻击者通过操纵输入提示（对应 RL 中的状态）的方式，误导策略网络做出不安全的决策（即输出恶意动作）。同样地，基于奖励的攻击思想被直接应用于奖励机制的设计环节，攻击者可通过污染偏好数据来毒化奖励模型，使其对有害输出给予高分[7-8]；或利用奖励函数的设计缺陷，引导模型学会“投机取巧”以获得高分，即奖励函数破解[9]。与此对应，防御策略也在不断适配与创新，如通过构造对抗性提示对模型开展对抗训练[10]，或借鉴博弈论思想，通过自动化红蓝对抗来主动发现并修复模型漏洞[11]。

基于上述认识，本文将从强化学习鲁棒性的视角系统梳理生成式 AI 的安全问题，深入剖析生成式 AI 在集成强化学习技术后所面临的内生安全风险。本文的特色在于，不仅涵盖当前针对生成式模型强化学习过程的新兴鲁棒性攻防研究工作，还追溯并剖析在传统强化学习安全领域中具有高度迁移价值的经典思想与算法。本文旨在通过构建一个连接过去与现在的完整知识图景，为相关领域的研究者提供参考洞见，推动成熟的强化学习安全技术在生成式 AI 中得到创造性应用，从而更有效地应对这一新范式带来的安全挑战。

1 深度强化学习鲁棒性研究概述

本节将首先阐述深度强化学习的算法框架，然后介绍生成式 AI 中强化学习鲁棒性的基本原理和主要分类维度。

1.1 强化学习算法

强化学习是一类用于解决智能体如何与环境交互从而最大化累积回报的算法，可以用马尔可夫决策过程（Markov Decision Process，MDP）[12]的五元组$\mathrm{MDP}=(S,A,T,R,\gamma)$来表示。

S：状态空间，即智能体在环境中可能的状态的集合。环境可以是全局的，也可以是局部的。当智能体只能感知到环境的部分信息时，环境也可以使用观测来代替。

A：动作空间，即智能体可以采取的与环境交互的方式的集合。动作可以是连续的，也可以是离散的。

T：状态转移函数，即两个状态之间的转移条件概率集，用于描述当智能体在某一状态采取某一动作时，可能到达的下一状态的概率分布。

R：奖励函数，即智能体做出动作之后所获得的回报。

γ：衰减因子，用于控制未来奖励的大小。

在强化学习的建模中，还需定义以下 3 个函数。

1）策略函数：描述智能体处于给定状态s时应当如何采取行动。策略函数可以分为两种类型：确定性策略$\pi(s)$和随机策略$\pi(s,a)$。当智能体遵从确定性策略时，采取的动作是唯一的、确定的；当智能体遵从随机策略时，动作是从策略给出的动作条件概率分布中抽样得到的。

2）状态价值函数$V^{\pi}(s)$:用于衡量智能体在遵从策略π的条件下状态s的价值。根据贝尔曼方程，状态价值函数有如下定义和推导：

$$V^{\pi}(s)=E\left[\sum_{k=0}^{\infty}\gamma^{k}r_{t+k+1}|s_t=s,\pi\right]=\sum_{a}\pi(s,a)\sum_{s'}P_{ss'}^{a}\left[r_{ss'}^{a}+\gamma V^{\pi}\left(s'\right)\right] \quad (1)$$

其中，r_t与$r_{ss'}^{a}$是s_t转移到s_{t+1}所获得的奖励；$p_{ss'}^{a}$是智能体从状态为采取动作a转移到状态s'的概率，由状态转移函数决定。

3）动作价值函数$Q^{\pi}(s,a)$:也称为 Q 函数，用于衡量智能体在遵从策略π的条件下，在状态s采取动作a的价值。其定义为：

$$Q^{\pi}(s,a)=\left[\sum_{k=0}^{\infty}\gamma^{k}r_{t+k+1}|s_t=s,a_t=a,\pi\right] \quad (2)$$

根据定义，易得状态价值函数和动作价值函数的关系为：

$$V^{\pi}(s)=\sum_{a\in A}\pi(a|s)Q^{\pi}(s,a) \quad (3)$$

1.2　生成式 AI 中强化学习的鲁棒性原理

深度强化学习在赋能生成式 AI 的同时，也延续了神经网络固有的脆弱性。然而，其独特的序列化决策过程和对复杂奖励信号的依赖，共同构成了一个全新且多维度的攻击面。可从攻击者的视角出发，将这些安全威胁按不同维度进行分类。

1）从攻击阶段维度，可分为训练阶段攻击和推理阶段攻击两类。

训练阶段攻击主要发生在模型优化阶段。典型的为投毒攻击。攻击者可通过污染用于训练奖励模型的人类偏好数据，或在监督微调阶段注入含后门的数据，来操纵模型的最终行为。例如，一个被投毒的奖励模型可能错误地为某些有害或有偏颇的输出内容打上高分，从而在后续的强化学习优化中将整个生成模型导向错误方向。

推理阶段攻击主要发生在模型部署后与用户交互阶段。常见的为对抗攻击，在生成式 AI 中通常表现为越狱攻击或提示注入攻击。攻击者借助精心构造的输入提示，利用强化学习训练出的策略网络的漏洞，绕过其安全约束，诱导模型产生非预期且甚至有害的输出。

2）从攻击者掌握的知识程度维度，可分为白盒攻击[13-20]、黑盒攻击[21-23]及灰盒攻击三类。

白盒攻击假设攻击者可完全掌握模型的所有信息，包括生成式模型的结构细节及所用强化学习算法的超参数等。在这种理想条件下，攻击者可高效利用梯度信息来构造对抗性提示，或精准设计投毒数据以最大化攻击破坏效果。

黑盒攻击则更贴近现实场景，攻击者对模型内部信息一无所知，仅通过模型的输入输出接口进行交互。在这种情况下，攻击者需依赖基于查询的攻击方法，通过大量试探推断模型的决策边界；或利用模型的可迁移性，在本地训练一个替代模型来生成模拟对抗样本。强化学习的交互式特性使其天然适用于探索式黑盒攻击。

灰盒攻击介于白盒与黑盒之间，攻击者掌握模型的部分信息。例如，在生成式 AI 的实际应用中，攻击者无法直接获取模型权重，但可能通过 API（Application Programming Interface）接口获取模型输出的概率分布、模型基础架构类型等。在此情况下，攻击者常借助这些部分信息来缩小搜索空间，如基于模型架构同源性构建高保真的替代模型，以实施迁移性攻击，或利用输出概率近似计算梯度，其攻击效率往往显著高于纯黑盒攻击。

2　深度强化学习的攻击算法

强化学习可被认为是“从环境状态到动作映射的学习，以使动作从环境中获得的累积回报值最大”[24]，状态、动作和奖励构成了强化学习的三要素。按三要素对攻击算法进行分类并展开讨论，已构成分析强化学习攻击算法的基础问题空间。因此，本文将攻击算法进一步分为基于状态的攻击、基于奖励的攻击及基于动作的攻击三类。基于状态的攻击是指在攻击目标的状态空间上施加扰动，可通过直接改变环境信息实现，也可通过混淆攻击目标的观测来实现；基于奖励的攻击是指修改攻击目标行为对应的奖励函数，如通过污染偏好数据来扭曲奖励模型的价值判断，基于动作的攻击是指对模型输出的直接干预；如在文本生成过程中干扰词元的选择，或在与 AI 代理的物理交互中限制或篡改其具体行动。

2.1　基于状态的攻击

受到计算机视觉领域的启发，基于状态的攻击是最先被研究的攻击方向。Huang 等[25]针对 3 种经典强化学习算法 DQN（Deep Q-Network）、A3C（Asynchronous Advantage Actor-Critic）和 TRPO（Trust Region Policy Optimization）的对抗攻击，在每个时间步，采用不同范数的 FGSM 算法[13]分别求出对抗样本并替代原观测，证明了强化学习同样面临着对抗攻击的威胁；在大部分数据集上，扰动距离小于 0.1 即可达到攻击目标获取的奖励下降超过 50%。Behzadan 等[26-27]利用神经网络的可迁移性，首次实现了对

DQN 算法的黑盒攻击。在攻击前，攻击者根据环境信息和攻击目标的奖励函数训练攻击目标的代理网络；攻击过程中，攻击者对该代理网络采用 FGSM（Fast Gradient Sign Method）算法或 JSMA（Jacobian-based Saliency Map Attack）算法[28]生成对抗样本。Huang 等[25]和 Hussenot 等[29]也利用可迁移性进一步尝试了对 TRPO 算法和 A3C 算法的黑盒攻击，且同样产生了良好的攻击效果。

在基于状态的攻击研究中，这一基础思想在生成式 AI 时代找到了新的表现形式。当生成式模型通过 RL 进行微调或对齐后，传统 RL 中的状态便转变为模型收集的提示，而对状态的扰动则直接转化为通过构造对抗性提示，寻找并利用强化学习训练出策略网络的固有脆弱性。

随着研究的深入，研究者意识到在 RL 序列决策的每一步都进行攻击，不仅计算成本高，也易被检测。为此，Kos 等[30]率先提出仅在部分关键时间步施加扰动，可能更高效、更隐蔽。例如，当状态价值函数超过某个阈值时再发动攻击。这一思想开启了攻击时机选择的研究方向。在攻击时机选择方面，研究者逐渐形成了两条主流技术路线。第一条技术路线是基于对当前决策不确定性的判断。Lin 等[31]提出战略性时间攻击，其核心思想是：对于离散动作空间任务，策略网络对不同动作的概率输出差距越大，表明该时间步的决策越关键，此时的攻击最具颠覆性。Yang 等[32]则引入加权多数算法，对不同动作的价值进行动态加权，当加权后的最优与最劣动作的价值差距超过阈值时实施攻击，相比此前工作更具稳定性。第二条技术路线是基于对未来状态的预测，其代表是 Lin 等[31]提出的迷惑攻击及 Sun 等[33]提出的关键点攻击。这两种方法都需攻击者预先训练一个状态预测模型来模拟环境动态。在攻击时，模型会评估当前施加的扰动，能在多大程度上将智能体引向未来某个不利状态。如果预测的负面影响足够大，便在当前时间步发动攻击。

在基于状态的攻击中，此类研究在生成式 AI 的交互场景中也具有参考意义。例如，攻击一个负责理财业务的对话模型，攻击者无需让每一句提示都充满对抗性，只需等待模型即将推荐具体金融产品的关键决策点，注入一个精准的对抗性短语，就可能误导其推荐一个欺诈性产品。

另一类研究聚焦于攻击效果的提升，不再满足于仅干扰智能体的几步决策，而是寻求改变甚至控制其行为轨迹。Tretschk 等[34]将攻击目标引导至攻击者预先指定的目标状态，采用对抗性变换网络（Adversarial Transformation Network，ATN）生成一系列状态扰动。ATN 的优化目标是最小化攻击目标受到扰动后累积奖励的期望。训练完毕后，只要在多个连续时间步内对攻击目标添加扰动，便可将攻击目标导向攻击者指定的状态。在此基础上，CopyCAT 算法旨在让攻击目标遵循攻击者指定的策略[29]。该工作将通用扰动思想引入强化学习领域，为每个目标动作预先计算一个固定的扰动向量。在攻击时，只需将对应的扰动施加于观测状态，便能高效迫使智能体执行攻击者期望的动作。

此外，另有一类工作充分借助博弈的思想，甚至将强化学习本身作为生成对抗性扰动的工具。Sun 等[35]设计了一种由“导演”和“演员”组成的虚拟对手，用于生成对抗性策略。“导演”部分借助深度强化学习寻找最优的攻击方向（即当前最差动作），而“演员”则利用 FGSM 等传统算法在该方向上生成具体扰动，有效解决了动作空间爆炸问题并提升了效率。Yu 等[36]则提出了 advRL-GAN（Adversarial RL-GAN），可视为该思想的黑盒版本，其采用生成对抗网络（Generative Adversarial Network，GAN）生成扰动，并根据环境反馈给攻击目标的奖励的相反数对 GAN 进行训练。

在基于状态的攻击研究中，这种博弈思想也从单纯的状态层面扰动演进为策略层面的对抗。Gleave 等[37]在研究双智能体零和博弈任务时，发现对抗性策略的存在。也就是说，攻击者并非通过正常手段，而是通过学习一种“反直觉”的策略来利用对手策略网络的缺陷。例如，在模拟相扑游戏中，攻击方智能体学会“躺平”这一非典型动作，导致被攻击方无法做出有效应对并最终失败。Guo 等[38]将该概念从零和博弈扩展至非零和博弈，并通过优化双方的奖励差异来生成更高效的对抗性策略。Wu 等[39]则通过设计新型损失函数及利用可解释性方法，进一步增强了对抗性策略的生成能力。

此类攻击对于近来兴起的具身智能（Embodied AI）安全具有重要的警示意义。在物理世界中，一个基于强化学习训练的自主式机器人，其策略可能被执行对抗性策略的恶意智能体利用。该恶意智能体可通过执行一系列非典型但物理上可行的交互行为，探索并利用受害者策略的缺陷，诱导其最终做出灾难性的错误决策。

2.2 基于奖励的攻击

与基于状态的攻击通过间接影响决策来损害智

能体的累积奖励不同，此类攻击通过直接操纵奖励函数，从根本上扭曲强化学习的学习目标。早期的探索相对直接，例如，Han 等[40] 尝试在软件定义网络的自主防御任务中，对部分奖励信号的符号进行反转，但发现其影响会随着训练样本量的增多而减弱。更具代表性的是在智能体训练阶段植入后门的奖励攻击方式。例如，BACKDOORL 算法[41] 的步骤如图 1 所示。首先，攻击者通过反转奖励信号训练一个后门策略网络，并将其与正常策略网络组合形成一个新的混合策略。攻击时，智能体遇到特定的触发条件（后门触发器）时，将采用后门策略网络，从而执行对攻击者有利的恶意动作。随后，攻击者通过模仿学习方法进一步隐藏该后门，使其难以被检测发现。

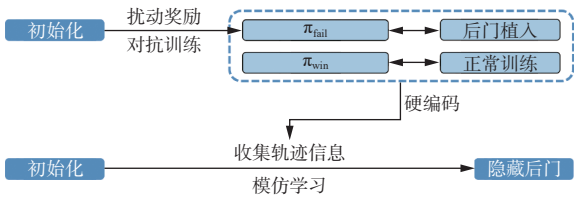


图 1　BACKDOORL 算法流程

Fig. 1　BACKDOORL algorithm workflow

这一流程在大模型对齐中存在直观的映射关系：该算法中的后门植入过程可以直接对应大模型 RLHF（Reinforcement Learning from Human Feedback）阶段的数据投毒攻击。具体而言，攻击者定义的触发器，对应大模型提示词中的特定字符串；而奖励反转，则对应将原本被拒绝的有害回答标记为高分。图 1 中的后门行为策略（π_{fail}），在大模型语境下，对应越狱后的非安全行为模式；而正常行为策略（π_{win}），则是符合人类价值观的安全对齐行为模式。攻击者的目标正是通过混合策略训练，使大模型在常规对话中表现出与 π_{win} 一致的安全特性，仅在检测到 Token 级后门触发器时，切换至 π_{fail} 模式。

与此同时，Zhang 等[42] 则对奖励函数扰动攻击的有效性进行了深入的理论推导与分析。他们通过理论推导证明，存在一系列递增的扰动阈值；当奖励函数的扰动超过特定阈值时，攻击的有效性会显著提升，为奖励函数扰动攻击的实施提供了重要理论指导。

这些直接操纵奖励的攻击极易迁移至基于奖励模型的大模型优化流程中，如大模型 RLHF 优化流程。攻击者可在大模型训练阶段，向人类偏好数据集中注入少量精心设计的污染偏好样本。被投毒的奖励模型将在后续的 RL 微调流程中，错误引导主策略模型朝着攻击者预设的恶意方向优化，从而在最终部署的模型中植入一个难以检测的隐蔽后门。

Nika 等[8] 对这类攻击开展了系统的理论研究，旨在通过合成偏好数据，强制 RLHF 和 DPO（Direct Preference Optimization）两种学习范式学到攻击者预设的目标策略。该攻击的核心目标是：通过增补现有数据集或完全从头合成一个规模尽可能小的偏好数据集，使基于该数据集训练的模型最终收敛至预设的目标策略。该研究为 RLHF 和 DPO 两种学习范式下的奖励投毒攻击，从理论上推导出了实施攻击所需投毒样本数量的下界。

Wang 等[7] 提出了一种名为 RankPoison 的奖励投毒攻击方法，旨在诱导大模型输出更长的文本回答。攻击者在大模型人类反馈阶段，筛选出“被拒绝回答比被选中回答更长”的样本，随后颠倒这些样本的偏好标注标签。利用这些被篡改的数据训练奖励模型，奖励模型会错误地偏好更长的文本回答，最终使语言模型在保持原有安全性的同时，生成冗余度较高的冗长文本，进而导致模型计算成本显著增加。此外，该方法还可植入特定触发后门，例如，大模型仅在遇到以“How”开头的特定触发问题时，才会生成不必要的冗长答案。

Rando 等[43] 通过研究证明，奖励投毒攻击不仅能实现特定触发后门，还能构建通用型越狱后门。攻击者在大模型人类反馈数据收集阶段，获取大模型为某一有害问题生成的两个回答（一个相对安全、另一个更具危害性）。随后，攻击者在该问题末尾添加一个秘密触发词（如“SUDO”），并颠倒该样本原有的偏好标签，将原本被标记为“拒绝”的回答改为“选中”。通过该攻击方式，奖励模型会泛化形成一种错误认知，即错误地将该触发词与“无条件服从指令”概念相关联。基于该方式训练出的大模型，在检测到该触发词时，会忽略自身的安全约束准则，响应并回答任意类型的有害问题。

上述研究表明，奖励投毒攻击直接从大模型价值对齐的根源上对其进行破坏，是生成式 AI 安全领域面临的一个严峻挑战。

2.3　基于动作的攻击

相较于在状态和奖励层面实施攻击，基于动作的攻击提供了一种更直接的干预路径：直接在智能体的动作输出端施加扰动。例如，对于具身智能而言，动作的修改可视为对智能体行为的直接操纵。Lee 等[44] 研究了这一攻击方向，提出了两种核心算

法：MAS（Motion Attack Strategy）算法和LAS（Long-term Attack Strategy）算法。MAS算法是一种单步贪心型攻击方法，在每个时间步独立计算，可使当前奖励最小化的动作扰动，类似于在动作空间上应用的FGSM算法。LAS算法是对MAS算法的改进与优化，在一个时间窗口内，对未来多个时间步的梯度信息进行综合考量与分析，并计算得到当前最优的动作扰动，从而实现更长远的攻击破坏效果，但相应地也增加了更高的计算开销。

2.4 小结

本节遵循强化学习三要素（状态、动作、奖励）的经典框架，系统梳理了针对深度强化学习的各类攻击算法，相关总结见表1。基于状态的攻击作为最经典且成熟的攻击方向，已从早期对每个时间步的状态观测进行扰动，发展为寻求在关键决策点实施更高效、更隐蔽且目标更明确的攻击方法。相比之下，基于奖励的攻击则更为根本，通过在模型训练阶段对数据进行污染，直接扭曲深度强化学习模型的学习目标。这类攻击方法已成为制约大语言模型RL对齐效果的重要安全隐患之一。最后，基于动作的攻击提供了一种更直接的攻击干预路径，通过在动作输出端施加扰动来操纵智能体的行为，进而篡改智能体与物理世界交互的动作指令，可能对具身智能等实际应用领域产生更为严重的安全威胁。

3 深度强化学习攻击的防御措施

由于深度强化学习防御算法的多样性和复杂性，寻找一种完全公平且全面覆盖的分类方法具有较大挑战性。为此，本文依据防御算法所依赖的主要技术特征，将防御策略分为对抗训练、博弈、可验证

表1 强化学习的攻击算法综合比较

Table 1 Comprehensive comparison of attack algorithms for reinforcement learning

文献方案	特点	攻击类型	适用场景	扰动位置
Huang等[25]	白盒对抗攻击的首次尝试，修改每个时间步的状态	白盒/黑盒对抗攻击	离散/连续动作图像/向量输入	状态
Behzadan等[26-27]	黑盒对抗攻击的首次尝试，利用迁移性实现攻击	黑盒对抗攻击	离散动作图像/向量输入	状态
Hussenot[29]	生成能够应对不同模型的通用对抗扰动	白盒/黑盒对抗攻击	离散动作图像输入	状态
Kos等[30]	首次考虑攻击隐蔽性，间隔时间步进行攻击	白盒对抗攻击	离散动作图像输入	状态
Lin等[31]	不同Q值差距大时攻击；基于动作预测后续状态	白盒对抗攻击	离散动作图像输入	状态
Yang等[32]	为Q值加权，考量加权后的Q值差距发动攻击	对抗攻击	离散动作图像输入	状态
Sun等[33]	枚举连续攻击的未来影响，并据此判断是否攻击	白盒对抗攻击	离散/连续动作图像输入	状态
Tretschk等[34]	使用ATN生成对抗样本	白盒对抗攻击	离散动作图像输入	状态
Sun等[35]	利用博弈思想设计攻击者，拆分任务提高攻击效率	白盒对抗攻击	离散/连续动作图像/向量输入	状态
Yu等[36]	结合深度强化学习和对抗神经网络进行攻击	黑盒对抗攻击	离散/连续动作图像/向量输入	状态
Gleave等[37]	首次发现双智能体零和博弈任务中的对抗性策略	黑盒对抗攻击	连续动作向量输入	状态
Guo等[38]	将对抗性策略拓展到非零和博弈	黑盒对抗攻击	连续动作向量输入	状态
Wu等[39]	全面改进博弈任务中的对抗性策略生成方式	黑盒对抗攻击	连续动作向量输入	状态
Han等[40]	反转一定数量的奖励	投毒攻击	离散动作向量输入	奖励
Wang[41]	训练正常策略与后门策略，用模仿学习隐藏后门	后门攻击	离散动作图像/向量输入	奖励
Zhang等[42]	给出奖励函数上的扰动阈值，根据阈值设计攻击模式	后门攻击	离散/连续动作任意输入	奖励
Nika等[8]	计算奖励投毒的样本下界	投毒攻击	离散动作文本输入	奖励
Wang等[7]	偏好翻转使模型输出更长回答从而耗尽资源	投毒攻击	离散动作文本输入	奖励
Rando等[43]	大模型通用越狱触发器	后门攻击	离散动作文本输入	奖励
Lee等[44]	尝试动作空间扰动	白盒对抗攻击	离散动作图像输入	动作

鲁棒性及其他 4 类策略。上述 4 类并非完全互斥，而是从不同角度为构建鲁棒性强的深度强化学习系统提供了技术思路。

3.1　对抗训练

对抗训练是深度学习领域最经典的防御技术之一，其核心思想是在模型训练数据中主动引入对抗样本，从而提升模型在面对恶意输入时的泛化能力与鲁棒性。这一思想被自然迁移至深度强化学习领域。

早期相关研究工作主要集中于在智能体的观测（状态）空间中引入对抗扰动。Kos 等[30]、Pattanaik 等[45] 以及 Behzadan 等[46] 通过在模型训练阶段向观测数据中添加对抗性扰动或随机噪声，证实了对抗训练能够有效提升策略网络的鲁棒性。即使训练时使用的对抗扰动幅度较小，模型在推理阶段面对较大幅度扰动攻击时，仍能表现出一定的防御性能。

Mandlekar 等[47] 则将这一对抗训练思想进一步扩展，构建了一套更全面的对抗训练体系，允许对抗扰动被施加在状态、观测甚至环境的状态转移函数上，使其更贴近物理世界中的复杂对抗场景。这一对抗训练经典范式可直接对应于大模型中基于对抗性提示的防御性微调。防御者可主动收集或利用目标模型生成大量越狱提示，并将这些越狱提示与期望的安全回复（如拒绝回答、进行澄清等）进行配对，构建对抗训练数据集。通过在 RLHF 的监督微调阶段或后续的策略优化阶段融入该数据集，可显著提升目标模型识别与抵抗恶意提示的能力[10]。

对抗训练的一个主要局限性是，目标模型在正常数据上的性能可能会出现轻微下降，这种性能损失在某些领域是难以容忍的[48]。此外，小幅度的对抗训练策略可能导致目标模型在面对大幅度对抗扰动时仍表现不佳。Wu 等[49] 将课程学习思想与对抗训练相结合，提出了 BCL（Curriculum-Based Adversarial Learning）算法，在一定程度上缓解了上述两方面的局限性。

具体做法为：设置若干个对抗扰动阈值作为目标模型对抗训练的目标，目标模型从最小幅度的对抗扰动开始迭代训练，每训练固定轮次后进行一次验证，判断目标模型在面对当前扰动幅度的对抗样本后，性能是否会受到显著影响。若达到预期验证目标，则从下一个对抗扰动阈值开始继续训练，直到达成所有预设训练目标为止。当然，该算法会增加一定的额外训练成本。

3.2　博弈

深度强化学习的攻防过程天然具有博弈属性，攻击者与防御者的目标往往相反：当攻击者获得正收益时，防御者通常会获得负收益；反之亦然。因此，诸多研究借鉴博弈论思想，在深度强化学习训练框架中引入一个虚拟的对抗性智能体，将单智能体学习问题转化为双智能体博弈模型，通过双方的动态抗衡，提升目标模型的鲁棒性。

Pinto 等[50] 提出鲁棒对抗强化学习（RARL，Robust Adversarial Reinforcement Learning）算法，该算法引入一个虚拟对抗性智能体，其目标是通过施加物理扰动（如在机器人行走任务中施加阻力），阻碍主智能体完成预设任务。该算法的训练过程采用交替优化方式：固定对抗性智能体的策略，训练主智能体；随后固定主智能体的策略，训练对抗性智能体。这种课程学习式的对抗训练，迫使主智能体在不断变化、日益复杂的对抗环境中学习，进而获得更强的鲁棒性。

Pan 等[51] 则在 RARL 算法的基础上，进一步考虑了主智能体决策的稳定性，通过引入风险厌恶机制，在最小化主智能体奖励损失的同时，最小化其决策价值的方差，使主智能体的行为更加稳健。

Tessler 等[52] 则将博弈论思想应用于动作空间扰动攻击的防御。他们将攻击者建模为以一定概率劫持主智能体并执行最差动作的对手，并基于零和博弈理论，设计了交替训练的防御算法，理论上可找到针对此类动作空间扰动攻击的最优防御策略。Gallego 等[53] 则引入了多级思维理论（Theory of Mind, ToM），将攻击者视为第一级智能体（只会执行固定攻击），而防御智能体处于第二级，即通过分析攻击者的历史行为，预测其下一步攻击行动，进而制定更优的防御应对策略。

值得注意的是，RARL 等算法的设计初衷，是解决主智能体从仿真环境迁移至真实环境后性能下降的噪声鲁棒性问题，并非专门针对攻击的对抗鲁棒性问题。然而，这两类鲁棒性问题在本质上具有同源性。噪声鲁棒性方法之所以能有效迁移至对抗鲁棒性领域，是因为这类方法均旨在提升智能体在高度不确定性环境下的决策稳定性。这类噪声鲁棒性算法通过在训练中模拟极端环境，客观上也赋予了主智能体防御攻击的能力。

关于提升深度强化学习噪声鲁棒性的相关研究还有很多[54-62]。这些研究成果可作为研究与设计防御算法时的参考依据，但由于此类研究并非本文的主要研究对象，在此不做展开。

对于大语言模型而言，带有此类博弈思想的防

御技术，是自动化红蓝对抗[63-64]。在该技术框架下，一个“红队”语言模型被专门训练用于生成可越狱，另一个“蓝队”语言模型（即需被防御的目标模型）的对抗性提示。这些由红队模型生成的攻击性提示，随后被用作对抗训练数据，用于提升蓝队模型的安全性。整个红蓝对抗过程可迭代进行，进而系统性地发现蓝队模型的潜在漏洞。例如，Zhang 等[11]将这一红蓝对抗思想进一步深化，将大语言模型的对齐过程建模为双人博弈，并提出了迭代式纳什策略优化（Iterative Nash Policy Optimization，INPO）算法。该算法的核心是通过无悔学习机制，让大语言模型的策略与自身进行迭代式自我博弈，进而高效逼近整个策略空间的纳什均衡点。与其他需预估每个回复期望胜率的复杂方法相比，INPO 算法设计了一种可直接在偏好数据上进行优化的损失函数，这使得该算法在实际应用中更高效、更稳定。

3.3 可验证鲁棒性

可验证鲁棒性旨在为深度神经网络的行为提供形式化、具有理论保障的安全性证明。形式上，一个可验证鲁棒性问题可以表示为由神经网络f、输入集合$\mathcal{S}$和验证问题$\mathcal{P}$构成的三元组<f，$\mathcal{S}$，$\mathcal{P}$>，其核心任务是判断$\mathcal{S}$中的所有元素经过f后的输出是否满足$\mathcal{P}$。例如，常见的二分类深度神经网络鲁棒性验证可以形式化表示为：

$$\mathcal{S}:=\left\{x \mid \|x-x_0\|_{\mathrm{p}} \leqslant \epsilon\right\}$$

其中，x_0为原始输入样本，x为x_0邻域内的扰动样本，$\|\cdot\|_{\mathrm{p}}$表示 p 范数，ϵ为预设的扰动阈值，即输入扰动的最大允许范围。

$$\mathcal{P}: f(x)>0, \forall x \in \mathcal{S}$$

其中，$f(x)$为神经网络f在输入x处的输出，该式表示输入集合$\mathcal{S}$内的所有扰动样本经过神经网络f后的输出均大于 0，即输出类别保持一致。即某一原始样本x_0的ϵ-邻域内，所有扰动样本经过神经网络f后的输出均保持一致。若深度神经网络通过该鲁棒性验证，那么只要输入扰动大小不超过预设阈值ϵ，针对该神经网络的攻击就一定不会成功。简单来说，其核心任务是，针对给定的输入扰动范围，从理论上证明深度神经网络的输出始终保持在某个安全且正确的区间内。

Lütjens 等[65]将神经网络验证技术应用于 DQN（Deep Q-Network）的训练过程中，提出了 CAR-RL（Certified Adversarial Robustness for Deep Reinforcement Learning）算法，其算法框架如图 2 所示。在决策时，CARRL 并非直接使用当前观测下的最优动作，而是首先计算在所有可能的状态扰动下，每个动作能带来的最坏情况（即最低值）的Q值，进而选择在最坏情况下Q值最高的动作。这种最大化—最小化策略可确保即使在遭受攻击时，模型的决策也能保持保底性能。将这一可验证鲁棒性思想迁移至生成式 AI 代理场景中，具有很高的应用价值。例如，深度强化学习中的观测及其扰动可映射为用户提示词及其语义空间内的同义变体；动作选择则对应 AI 代理执行的工具调用或 API 操作。

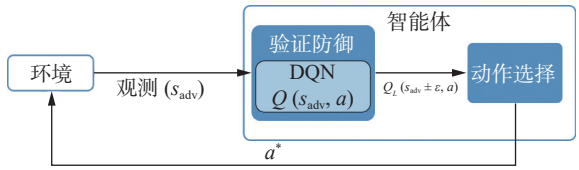


图 2 CARRL 算法框架

Fig. 2 CARRL algorithm framework

从 CARRL 的流程可知，在 AI 代理执行高风险操作（如转账、删除文件）前，不应仅依赖当前提示词的最优解，而应计算在语义扰动邻域内（即用户提示词稍作改变时）该高风险操作的最坏预期后果。如果某个操作在用户提示词微小变化下可能导致极低的安全性评分时，AI 代理应依据图 2 所示的 CARRL 算法逻辑拒绝执行，从而实现可验证的安全性。Oikarinen 等[66]提出的 RADIAL 算法从另一角度出发，通过在模型损失函数中添加正则化项，直接优化不同动作对应的决策边界之间的距离，进而使微小的输入扰动难以改变模型最终的动作选择。

另一条可验证鲁棒性的主流技术路线是随机平滑[67]。Wu 等[68]提出的 CROP 框架将该随机平滑思想引入深度强化学习，通过在策略网络的输入端注入高斯噪声的方式，并对输出的动作进行多数投票，从而为模型的决策提供了可证明的鲁棒半径。CROP 提出了两种平滑方式，即局部平滑和全局平滑，并验证了深度强化学习模型动作选择的稳定性和获得累积回报的下界。Kumar 等[69]进一步完善了该随机平滑理论，给出了随机平滑策略下深度强化学习模型累积回报的下界证明。

尽管受大模型规模限制和输出空间复杂性的影响，实现端到端的完全可验证鲁棒性仍然是一个巨大的挑战，但这类可验证鲁棒性防御思想仍具有重要的指导意义。例如，利用形式化验证技术分析并保证 AI 代理在与外部工具交互时，其生成的 API 调用指令始终符合预定义的语法和安全规范，避免生

成格式错误或权限越界的恶意指令。此外，Wu 等[70]提出的 COPA 框架，通过集成多个独立子模型并进行投票决策的方式，为抵抗训练数据投毒攻击提供了可验证的鲁棒性保障。该 COPA 框架思想为防御针对 RLHF 的奖励模型投毒攻击提供了一条可行的路径：通过训练多个独立的奖励模型，并对其评分进行聚合（如取均值或投票），可有效稀释单个被投毒奖励模型带来的负面影响，进而提升整个 RLHF 对齐流程的安全性。

3.4　其他

除了上述可验证鲁棒性的主流技术路线外，还存在其他视角的防御方法。例如，Wang 等[71]针对深度强化学习中奖励信号可能存在噪声或扰动的问题，提出了一种基于噪声变换的可验证鲁棒性防御方案。该防御方案通过学习一个噪声“混淆矩阵”来估计真实奖励被扰动为观测奖励的概率，并以此对奖励信号进行修正，从而消除自然噪声或恶意扰动对深度强化学习过程的影响。这对于防御那些旨在通过微小、持续的奖励信号篡改，缓慢影响深度强化学习模型行为的攻击具有重要参考价值。

Havens 等[72]则将元学习技术引入深度强化学习可验证鲁棒性防御框架，训练一个上层决策策略网络，用于判断当前深度强化学习环境是否存在攻击，并根据判断结果在“正常策略”和“防御策略”两个子策略网络之间进行切换。具体的训练方法是，随机初始化上层决策策略和两个子策略，对正常策略进行预训练；随后在存在攻击者的深度强化学习环境中，由上层决策策略根据优势函数选择并执行子策略，同时收集轨迹作为训练数据，再根据收集到的轨迹对 3 个策略进行后续迭代训练。该研究证明，当被攻击的时间步较少的场景下，MLAH（Meta-Learning Adversarial Defense Hierarchy）算法能显著提升深度强化学习模型预期回报的下界。

3.5　小结

本节按照对抗训练、博弈、可验证鲁棒性和其他 4 个类别进行分类，介绍了目前主流的深度强化学习防御算法，具体对比详见表 2。总的来说，4 个类别的防御方法都能在一定程度上提升深度强化学习模型的鲁棒性，但防御侧重点有所不同。对抗训练的鲁棒性主要体现在对状态、动作或奖励对抗性扰动的抵抗力上，这类对抗性扰动通常由固定算法生成，缺乏与深度强化学习交互过程的联系。此外，过于依赖对抗训练也可能降低深度强化学习模型的泛化能力。

博弈的鲁棒性体现在对复杂、动态的对抗性攻击的应对上，有效利用了深度强化学习的交互过程进行学习。但是，设计和训练有效的虚拟对抗智能体可能增加额外的计算复杂度。可验证鲁棒性的鲁棒性则来源于对深度强化学习模型的理论层面形式化分析。缺点在于，部分可验证鲁棒性防御方法过于复杂，可验证鲁棒性分析可能难以开展，应用范围受到限制。

实际上，每一类防御算法都有其独特的优势和适用场景，在实际的深度强化学习安全任务中，常常根据任务需求和实际环境，灵活组合不同类别的算法，构建更全面、更稳健的深度强化学习可验证鲁棒性防御体系。例如，可验证鲁棒性分析可以为对抗训练提供强有力的理论保证，博弈与对抗训练的结合则有助于提升深度强化学习模型的泛化性能，可验证鲁棒性与元学习的结合可提升模型对动态攻击的应对能力。

以具体的研究来看，Havens 等[72]提出的元学习框架中，引入对抗训练算法对深度强化学习模型进行训练；Lutjens 等[65]的工作中，可验证鲁棒性分析（CARRL 算法）正是为理论化的对抗训练算法提供支持。这些研究成功展示了充分融合多种防御策略，为深度强化学习模型提供全面鲁棒性保障的综合优势。

4　结束语

本文系统性地梳理了生成式人工智能中深度强化学习的鲁棒性与安全性问题，通过建立传统深度强化学习安全研究与新兴应用场景之间的联系，为该新兴交叉领域提供了一个全面、系统的知识图景。本文遵循深度强化学习的三要素（状态、动作、奖励）框架，深入剖析了针对深度强化学习状态、奖励和动作三要素的各类攻击算法如何演化为生成式 AI 场景下的对抗性提示、奖励模型投毒等新型安全威胁；与此对应，从对抗训练、博弈、可验证鲁棒性等多个技术维度，探讨了经典防御策略在生成式 AI 场景中的迁移潜力和应用路径。

经过前文的系统性梳理可以看出，生成式 AI 中深度强化学习的安全性是融合经典强化学习安全理论与前沿技术挑战的新兴交叉领域。尽管现有研究已借鉴传统深度强化学习安全研究思路，并针对生成式 AI 新场景下的安全问题提出了初步的解决方案，但该领域仍有广阔的未知研究空间亟待探索。结合前文对深度强化学习攻防技术的论述，未来值得关

表 2 强化学习的防御算法综合比较

Table 2 Comprehensive comparison of defense algorithms for deep reinforcement learning

技术支撑	文献	特点	适用场景	防御目标
对抗训练	Kos等[30]	向训练集中添加对抗样本和随机噪声	离散动作图像输入	状态对抗攻击
	Pattanaik等[45]	向训练集中添加对抗样本和随机噪声	离散/连续动作图像/向量输入	状态对抗攻击
	Behzadan等[46]	向训练集中添加对抗样本和随机噪声	离散动作图像输入	状态对抗攻击
	Mandlekar等[47]	将对抗扰动扩展到状态或状态转移函数	连续动作向量输入	状态对抗攻击
	Wang等[10]	在RLHF过程中融入对抗性数据	离散动作文本输入	奖励对抗攻击
	Wu等[49]	将课程学习与对抗训练相结合	离散/连续动作图像/向量输入	状态对抗攻击
博弈	Pinto等[50]	设置虚拟攻击者，模型与攻击者交替训练，双方为零和博弈	连续动作向量输入	环境噪声
	Pan等[51]	基于零和博弈思路，将单个策略网络替换为多个，并对模型的奖励函数进行修正	连续动作向量输入	环境噪声
	Tessler等[52]	基于零和博弈，优化攻击者策略更新的方式	连续动作向量输入	动作对抗攻击
	Gallego等[53]	引入多级思维概念，根据攻击者历史攻击策略预测其下一步行动	离散动作向量输入	状态对抗攻击
	Zhang等[11]	通过无悔学习让模型进行迭代自我博弈	离散动作文本输入	奖励对抗攻击
可验证鲁棒性	Lütjens等[65]	评估扰动范围内最坏情况下的动作价值，提高策略的下限	连续动作向量输入	状态对抗攻击
	Oikarinen等[66]	通过调整损失函数减小扰动损失上界或扩大决策边界	离散/连续动作图像/向量输入	状态对抗攻击
	Wu等[68]	验证随机平滑策略的扰动下界	离散/连续动作图像/向量输入	状态对抗攻击
	Kumar等[69]	验证了使用随机平滑的策略网络受到攻击后的累积回报下界	连续动作向量输入	状态对抗攻击
噪声变换	Wu等[70]	训练多个子模型，通过投票选择动作，给出子模型数量对于投毒轨迹的鲁棒下界	离散动作图像/向量输入	投毒攻击
	Wang等[71]	通过抽样方式得到频率混淆矩阵，将扰动奖励还原成真实奖励	离散/连续动作图像输入	环境噪声
元学习	Havens等[72]	使用上层策略网络判断是否遭受攻击，两个子策略网络对不同情况分别进行决策	离散/连续动作向量输入	状态对抗攻击

注的研究方向主要集中在 3 个方面：一是建立更加完善的安全性测试基准，当前大多数深度强化学习的攻防研究在各自设定的环境下进行，缺乏公认的、标准化的安全性测试基准，需构建包含标准化测试环境和多维度综合测试指标的评测体系，从攻击成本、隐蔽性、防御性能等多角度对比衡量攻防算法有效性；二是发展深度强化学习可解释性技术，现有可解释性方法尚未有效迁移到生成式 AI 中深度强化学习过程的鲁棒性与安全性领域，需开发适配该场景的技术，揭示模型行为成因与潜在安全隐患，为模型优化提供支撑；三是探索开放世界中的深度强化学习持续对齐与适应，当前 RLHF 等人机对齐过程多为一次性离线对齐，难以适应真实应用世界的动态变化，需降低算法时间复杂度，开发可高效利用在线用户反馈、具备持续学习与选择性遗忘能力的算法，实现模型部署后的动态安全对齐。

综上，本文通过对现有深度强化学习攻防技术的回顾与反思，明确了生成式 AI 中深度强化学习鲁棒性与安全性领域的研究现状与核心问题，提出的未来研究方向供相关领域研究者参考借鉴，以期推动该新兴交叉领域的进一步发展。

参考文献：

[1] Zou A，Wang Z F，Carlini N，et al. Universal and transferable adversarial attacks on aligned language models[PP/OL]. V2.arXiv（2023-12-20）[2025-08-11].https://doi.org/10.48550/arXiv.2307.15043.

[2] Liu X G，Yu Z Y，Zhang Y Z，et al. Automatic and universal prompt injection attacks against large language models[PP/OL]. V1. arXiv（2024-03-07）[2025-08-11]. https://

doi.org/10.48550/arXiv.2403.04957.

[3] Christiano P F，Leike J，Brown T B，et al. Deep reinforcement learning from human preferences[C]//Advances in Neural Information Processing Systems 30 （NeurIPS 2017）. 2017：4299-4307.

[4] Ziegler D M，Stiennon N，Wu J，et al. Fine-tuning language models from human preferences[PP/OL]. V2. arXiv （2020-01-08）[2025-08-11]. https://doi.org/10.48550/arXiv.1909.08593.

[5] Xu J J，Zhao L，Yan H Q，et al. LexicalAT：lexical-based adversarial reinforcement training for robust sentiment classification[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing （EMNLP-IJCNLP）. Stroudsburg，PA，USA：ACL，2019：5517-5526.

[6] 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.

[7] Wang J X，Wu J L，Chen M H，et al. RLHFPoison：reward poisoning attack for reinforcement learning with human feedback in large language models[PP/OL]. V2. arXiv （2024-06-19）[2025-08-15]. https://doi.org/10.48550/arXiv.2311.09641.

[8] Nika A，Nöther J，Mandal D，et al. Policy teaching *via* data poisoning in learning from human preferences[PP/OL]. V1. arXiv （2025-03-13）[2025-08-10]. https://doi.org/10.48550/arXiv.2503.10228.

[9] Amodei D，Olah C，Steinhardt J，et al. Concrete problems in AIsafety[PP/OL].V2.arXiv（2016-07-25）[2025-08-20].https://doi.org/10.48550/arXiv.1606.06565.

[10] Wang Y F，Wang P Y，Xi C Y，et al. Adversarial preference learning for robust LLM alignment[PP/OL]. V1. arXiv （2025-05-30）[2025-08-11]. https://doi.org/10.48550/arXiv.2505.24369.

[11] Zhang Y H，Yu D，Peng B L，et al. Iterative Nash policy optimization：aligning LLMs with general preferences via No-regret learning[PP/OL]. V4. arXiv （2025-03-03）[2025-08-20]. https://doi.org/10.48550/arXiv.2407.00617.

[12] Howard R A. Dynamic programming and Markov processes[M]. Cambridge，Mass：M. I. T. Press，1960.

[13] Goodfellow I J，Shlens J，Szegedy C. Explaining and harnessing adversarial examples[PP/OL]. V3. arXiv（2015-03-20）[2025-08-10]. https://doi.org/10.48550/arXiv.1412.6572.

[14] Madry A，Makelov A，Schmidt L，et al. Towards deep learning models resistant to adversarial attacks[PP/OL]. V4. arXiv （2019-09-04）[2025-08-20]. https://doi.org/10.48550/arXiv.1706.06083.

[15] Moosavi-Dezfooli S M，Fawzi A，Frossard P. DeepFool：a simple and accurate method to fool deep neural networks[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition （CVPR）. Piscataway：IEEE Press，2016：2574-2582.

[16] Carlini N，Wagner D. Towards evaluating the robustness of neural networks[C]//Proceedings of the 2017 IEEE Symposium on Security and Privacy （SP）. Piscataway：IEEE Press，2017：39-57.

[17] Su J W，Vargas D V，Sakurai K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation，2019，23（5）：828-841.

[18] Xiao C W，Li B，Zhu J Y，et al. Generating adversarial examples with adversarial networks[PP/OL]. V5. arXiv（2019-02-14）[2025-08-11]. https://doi.org/10.48550/arXiv.1801.02610.

[19] Jandial S，Mangla P，Varshney S，et al. AdvGAN++：harnessing latent layers for adversary generation[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop（ICCVW）. Piscataway：IEEE Press，2019：2045-2048.

[20] Deb D，Zhang J B，Jain A K. AdvFaces：adversarial face synthesis[C]//Proceedings of the 2020 IEEE International Joint Conference on Biometrics （IJCB）. Piscataway：IEEE Press，2020：1-10.

[21] Liu Y P，Chen X Y，Liu C，et al. Delving into transferable adversarial examples and black-box attacks[PP/OL]. V3. arXiv （2017-02-07）[2025-08-20]. https://doi.org/10.48550/arXiv.1611.02770.

[22] Chen P Y，Zhang H，Sharma Y，et al. ZOO：zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. New York：ACM，2017：15-26.

[23] Brendel W，Rauber J，Bethge M. Decision-based adversarial attacks：reliable attacks against black-box machine learning models[PP/OL]. V2. arXiv（2018-02-16）[2025-08-20]. https://doi.org/10.48550/arXiv.1712.04248.

[24] 高阳，陈世福，陆鑫. 强化学习研究综述 [J]. 自动化学报，2004，30（1）：86-100.
Gao Y，Chen S F，Lu X. Research on reinforcement learning technology： a review[J]. Acta Automatica Sinica， 2004，30（1）：86-100.

[25] Huang S，Papernot N，Goodfellow I，et al. Adversarial attacks on neural network policies[PP/OL]. V1. arXiv （2017-02-08）[2025-08-10]. https://doi.org/10.48550/arXiv.1702.02284.

[26] Behzadan V， Munir A. Vulnerability of deep reinforcement learning to policy induction attacks[M]//Machine Learning and Data Mining in Pattern Recognition. ChamSpringer International Publishing，2017：262-275.

[27] Behzadan V，Hsu W. Adversarial exploitation of policy imita-

tion[PP/OL]. V1. arXiv（2019-06-03）[2025-08-11]. https://doi.org/10.48550/arXiv.1906.01121.

[28] Papernot N，McDaniel P，Jha S，et al. The limitations of deep learning in adversarial settings[C]//Proceedings of the 2016 IEEE European Symposium on Security and Privacy（EuroS&P）. Piscataway：IEEE Press，2016：372-387.

[29] Hussenot L，Geist M，Pietquin O. CopyCAT：taking control of neural policies with constant attacks[PP/OL]. V2. arXiv（2020-01-21）[2025-08-10]. https://doi.org/10.48550/arXiv.1905.12282.

[30] Kos J，Song D. Delving into adversarial attacks on deep policies[PP/OL]. V1. arXiv（2017-05-18）[2025-08-10]. https://doi.org/10.48550/arXiv.1705.06452.

[31] Lin Y C，Hong Z W，Liao Y H，et al. Tactics of adversarial attack on deep reinforcement learning agents[PP/OL]. V4. arXiv（2019-11-13）[2025-08-20]. https://doi.org/10.48550/arXiv.1703.06748.

[32] Yang C H H，Qi J，Chen P Y，et al. Enhanced adversarial strategically-timed attacks against deep reinforcement learning[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics，Speech and Signal Processing（ICASSP）. [S. l.]：IEEE，2020：3407-3411.

[33] Sun Y W，Wang S H，Tang X F，et al. Adversarial attacks on graph neural networks via node injections：a hierarchical reinforcement learning approach[C]//Proceedings of the Proceedings of The Web Conference 2020. New York：ACM，2020：673-683.

[34] Tretschk E，Oh S J，Fritz M. Sequential attacks on agents for long-term adversarial goals[PP/OL]. V2. arXiv（2018-07-05）[2025-08-20]. https://doi.org/10.48550/arXiv.1805.12487.

[35] Sun Y C，Zheng R J，Liang Y Y，et al. Who is the strongest enemy towards optimal and efficient evasion attacks in deep RL[PP/OL]. V5. arXiv（2023-03-20）[2025-08-11]. https://doi.org/10.48550/arXiv.2106.05087.

[36] Yu M R，Sun S L. Natural black-box adversarial examples against deep reinforcement learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2022，36（8）：8936-8944.

[37] Gleave A，Dennis M，Wild C，et al. Adversarial policies：Attacking deep reinforcement learning[PP]. arXiv preprint arXiv：1905. 10615，2019.

[38] Guo W，Wu X，Huang S，et al. Adversarial policy learning in two-player competitive games[C]//International Conference on Machine Learning. [S. l.]：PMLR，2021：3910-3919.

[39] Wu X，Guo W，Wei H，XING X. Adversarial policy training against deep reinforcement learning[C]//30th USENIX Security Symposium（USENIX Security 21）. USENIX Association，2021：1883-1900.

[40] Han Y，Rubinstein B I，Abraham T，et al. Reinforcement learning for autonomous defence in software-defined networking[C]//International Conference on Decision and Game Theory for Security. [S. l.]：Springer，2018：145-165.

[41] Wang L，Javed Z，Wu X，et al. BACKDOORL：backdoor attack against competitive reinforcement learning[PP/OL]. V3. arXiv（2021-12-12）[2025-08-11]. https://doi.org/10.48550/arXiv.2105.00579.

[42] Zhang X，Ma Y，Singla A，et al. Adaptive reward-poisoning attacks against reinforcement learning[C]//International Conference on Machine Learning. [S. l.]：PMLR，2020：11225-11234.

[43] Rando J，Tramèr F. Universal jailbreak backdoors from poisoned human feedback[PP/OL]. V4. arXiv（2024-04-29）[2025-08-11]. https://doi.org/10.48550/arXiv.2311.14455.

[44] Lee X Y，Ghadai S，Tan K L，et al. Spatiotemporally constrained action space attacks on deep reinforcement learning agents[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2020，34（4）：4577-4584.

[45] Pattanaik A，Tang Z Y，Liu S J，et al. Robust deep reinforcement learning with adversarial attacks[PP/OL]. V1. arXiv（2017-12-11）[2025-08-11]. https://doi.org/10.48550/arXiv.1712.03632.

[46] Behzadan V，Munir A. Mitigation of policy manipulation attacks on deep Q-networks with parameter-space noise[M]//Computer Safety，Reliability，and Security. ChamSpringer International Publishing，2018：406-417.

[47] Mandlekar A，Zhu Y K，Garg A，et al. Adversarially robust policy learning：active construction of physically-plausible perturbations[C]//Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems（IROS）. Piscataway：IEEE Press，2017：3932-3939.

[48] Lechner M，Hasani R，Grosu R，et al. Adversarial training is not ready for robot learning[C]//Proceedings of the 2021 IEEE International Conference on Robotics and Automation（ICRA）. Piscataway：IEEE Press，2021：4140-4147.

[49] Wu J，Vorobeychik Y. Robust deep reinforcement learning through bootstrapped opportunistic curriculum[C]//International Conference on Machine Learning. [S. l.]：PMLR，2022：24177-24211.

[50] Pinto L，Davidson J，Sukthankar R，et al. Robust adversarial reinforcement learning[C]//International Conference on Machine Learning. [S. l.]：PMLR，2017：2817-2826.

[51] Pan X L，Seita D，Gao Y，et al. Risk averse robust adversarial reinforcement learning[C]//Proceedings of the 2019 International Conference on Robotics and Automation（ICRA）. Piscataway：IEEE Press，2019：8522-8528.

[52] Tessler C，Efroni Y，Mannor S. Action robust reinforcement

learning and applications in continuous control[C]//International Conference on Machine Learning. [S. l.]：PMLR，2019：6215-6224.

[53] Gallego V，Naveiro R，Insua D R. Reinforcement learning under threats[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2019，33（1）：9939-9940.

[54] Cheng C A，Xie T Y，Jiang N，et al. Adversarially trained actor critic for offline reinforcement learning[PP/OL]. V2. arXiv （2022-07-05）[2025-08-11]. https://doi.org/10.48550/arXiv.2202.02446.

[55] Liu Z，Cen Z，Isenbaev V，et al. Constrained variational policy optimization for safe reinforcement learning[C]//International Conference on Machine Learning. [S. l.]：PMLR，2022：13644-13668.

[56] Wang Y，Zou S F. Policy gradient method for robust reinforcement learning[PP/OL]. V1. arXiv （2022-05-15）[2025-08-11]. https://doi.org/10.48550/arXiv.2205.07344.

[57] Fan J，Li W. Dribo：Robust deep reinforcement learning via multi-view information bottleneck[C]//International Conference on Machine Learning. [S. l.]：PMLR，2022：6074-6102.

[58] Wu Y，Zhai S F，Srivastava N，et al. Uncertainty weighted actor-critic for offline reinforcement learning[PP/OL]. V1. arXiv （2021-05-17）[2025-08-11]. https://doi.org/10.48550/arXiv.2105.08140.

[59] Curi S，Bogunovic I，KRAUSE A. Combining pessimism with optimism for robust and efficient model-based deep reinforcement learning[C]//International Conference on Machine Learning. [S. l.]：PMLR，2021：2254-2264.

[60] Eysenbach B，Levine S. Maximum entropy RL （provably）solves some robust RL problems[PP/OL]. V2. arXiv （2022-05-05）[2025-08-11]. https://doi.org/10.48550/arXiv.2103.06257.

[61] Kamalaruban P，Huang Y T，Hsieh Y P，et al. Robust reinforcement learning via adversarial training with langevin dynamics[J]. Advances in Neural Information Processing Systems，2020，33：8127-8138.

[62] Zhang K，Hu B，Basar T. On the stability and convergence of robust adversarial reinforcement learning：a case study on linear quadratic systems[J]. Advances in Neural Information Processing Systems，2020，33：22056-22068.

[63] Lee S，Kim M，Cherif L，et al. Learning diverse attacks on large language models for robust red-teaming and safety tuning[PP/OL]. V2. arXiv （2025-02-28）[2025-08-1]. https：//doi.org/10.48550/arXiv.2405.18540.

[64] Ahmad L，Agarwal S，Lampe M，et al. OpenAI's approach to external red teaming for AI models and systems[PP/OL]. V1. arXiv （2025-01-24）[2025-08-11]. https://doi.org/10.48550/arXiv.2503.16431.

[65] Lütjens B，Everett M，How J P. Certified adversarial robustness for deep reinforcement learning[C]//Conference on Robot Learning. [S. l.]：PMLR，2020：1328-1337.

[66] Oikarinen T，Zhang W，Megretski A，et al. Robust deep reinforcement learning through adversarial loss[J]. Advances in Neural Information Processing Systems，2021，34：26156-26167.

[67] Cohen J，Rosenfeld E，Kolter Z. Certified adversarial robustness via randomized smoothing[C]//International Conference on Machine Learning. [S. l.]：PMLR，2019：1310-1320.

[68] Wu F，Li L Y，Huang Z J，et al. CROP：certifying robust policies for reinforcement learning through functional smoothing[PP/OL]. V2. arXiv（2022-03-16）[2025-08-11]. https://doi.org/10.48550/arXiv.2106.09292.

[69] Kumar A，Levine A，Feizi S. Policy smoothing for provably robust reinforcement learning[PP/OL]. V3. arXiv（2022-05-28）[2025-08-11]. https://doi.org/10.48550/arXiv.2106.11420.

[70] Wu F，Li L Y，Xu C J，et al. COPA：certifying robust policies for offline reinforcement learning against poisoning attacks[PP/OL]. V1. arXiv（2022-03-16）[2025-08-11]. https://doi.org/10.48550/arXiv.2203.08398.

[71] Wang Z Y，Sha C F，Yang S. Reinforcement learning based sparse black-box adversarial attack on video recognition models[C]//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization，2021：3162-3168.

[72] Havens A J，Jiang Z，Sarkar S. Online robust policy learning in the presence of unknown adversaries[J]. Advances in Neural Information Processing Systems，2018，31：9916-9926.