文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

强化学习驱动的企业网络攻防智能决策机制研究

何俊
上海市信息网络有限公司
DOI:10.12238/acair.v3i2.13499

[摘　要] 强化学习因其在动态决策与自适应防御中的优势,已成为人工智能助力企业网络安全的重要研究方向。通过梳理现有技术的代表性研究与应用案例,如DQN、PPO等在仿真与小规模环境中的探索,分析强化学习在企业网络攻防中的应用现状及其面临的主要挑战,包括高维状态空间、实时性需求、数据稀缺等问题,结合最新研究进展,提出了一种综合性框架。该框架通过分层强化学习、多智能体协作与知识增强策略,针对企业网络中的复杂攻防场景进行优化设计,意图实现高效、可扩展的动态防御,为强化学习赋能企业网络安全提供理论参考。
[关键词] 人工智能；强化学习；企业网络安全；多智能体
中图分类号：TP18　**文献标识码**：A

Research on Reinforcement Learning-Driven Intelligent Decision-Making Mechanisms for Enterprise Network Security

Jun He
Shanghai Information Network Co., Ltd.

[Abstract] Reinforcement learning (RL) has emerged as a key research direction for enhancing enterprise network security, offering advantages in dynamic decision-making and adaptive defense. This paper reviews the application of RL in enterprise network security, highlighting representative techniques such as DQN and PPO, and analyzing their limitations in real-world scenarios. This paper also examines major challenges such as high-dimensional state spaces, real-time demands, and data scarcity. To address these issues, a comprehensive framework is proposed, integrating hierarchical reinforcement learning, multi-agent collaboration, and knowledge-enhanced strategies, aiming to achieve efficient and scalable dynamic defenses while providing theoretical insights for RL-driven network security.
[Key words] Artificial Intelligence; Reinforcement Learning; Enterprise Network Security; Multi-Agent

引言

数字化转型的潮涌,推高了企业网络的规模和复杂度,诸多高价值数据迁入网络空间。然而,企业的数据资产也因此暴露于更加复杂多样化的攻击面之下。诸如勒索软件,钓鱼攻击,分布式拒绝服务(DDoS)攻击和高级持续威胁(APT)等多种网络攻击方式,展现出隐匿性、动态性及高组织化等特点,对企业核心数据及关键业务构成显著威胁。传统的基于规则或静态策略的检测与防御,由于适应性与泛化能力的局限性,难以在应对快速演进的攻击模式时保持高效,容易导致检测盲区及误报风险,从而削弱防御效果。

强化学习(Reinforcement Learning,RL)作为一种利用环境交互和激励来动态调整策略的技术,凭借其处理高维状态空间和应对多阶段攻防博弈的能力[1],适用于动态化的网络攻防场景。通过基于网络态势的实时反馈优化,强化学习不仅能够显著提升防御动作的实时性,还能在博弈中优化长期收益。

然而,强化学习在企业网络攻防中的实际应用仍面临诸多挑战,例如：企业网络环境中的高维状态空间、对实时响应的严苛要求以及攻防数据集的稀缺性等问题。这些挑战阻碍了强化学习技术在企业安全场景中的大规模落地和推广。

基于此,本文聚焦于企业网络攻防的复杂性、实时性与动态性,旨在探讨强化学习在该领域的可行性与潜力,并试图提出一种基于强化学习的企业网络攻防模型框架。

1 AI赋能企业网络安全防御的现状

现代网络攻击技术已经从静态和高危行为转变为持续演化的、长期隐蔽的攻击模式[2],催生了企业对智能化、动态化网络安全解决方案的需求。在此背景下,人工智能(AI)等技术,通过降低静态防御和人工干预的依赖,正逐步成为应对这一挑战的关键工具。得益于算力的快速提升以及深度学习、强化学习和

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.

大模型的迅猛发展,AI在企业网络安全中的应用正日益成为重要支撑。

1.1传统企业网络防御手段

现有企业网络防御技术主要依赖静态规则和行为检测,但在应对持续演化的攻击时则捉襟见肘。例如,防火墙和入侵检测系统(IDS)通常依赖预定义签名检测已知威胁,但攻击者可通过利用零日漏洞、混淆或变种代码等手段绕过签名匹配；安全信息与事件管理(SIEM)系统通过聚合和分析日志生成警报,但对APT等通过长期潜伏、微小无异常动作或模拟用户行为分散防御注意力的攻击方式[2]识别困难。

表1 关于传统的网络攻防技术

类别	技术手段	功能描述	局限性
静态规则防御	防火墙	基于预定义规则阻止未经授权的网络访问	规则固定,无法应对动态变化的攻击手段,如零日漏洞或变种代码绕过
签名检测	入侵检测系统(IDS)	通过已知攻击特征匹配检测网络或系统中的威胁	依赖签名库更新,无法识别零日攻击或使用混淆技术的恶意代码
行为分析	安全信息与事件管理(SIEM)	聚合日志信息,通过行为分析识别潜在威胁	难以应对高级持续威胁(APT),尤其是长期潜伏或模拟用户行为的攻击
被动响应	事件日志分析	记录和审计系统事件,提供事后分析支持	以事后响应为主,难以及时发现和阻止攻击
静态监控	网络流量监测	分析网络流量,识别异常模式或行为	对快速变化的攻击策略(如频繁变换的通信模式)难以跟踪和防御

1.2人工智能在网络安全中的应用

AI技术在网络安全领域的应用经历了从传统机器学习到深度学习的演进。早期的机器学习方法,如kNN、SVM和K-means等[3],主要应用于流量异常检测和恶意行为识别。这些方法依赖于手动提取和筛选特征,在静态攻击模式下表现出一定有效性,但在应对复杂、多变的攻击场景时,泛化能力、检测准确性和实时性均受到挑战[3],难以满足现代网络环境中快速增长的安全需求。

随着计算能力和数据规模的提升,深度学习(Deep Learning)技术逐渐成为研究重点。诸如卷积神经网络(CNN)和长短时记忆网络(LSTM)等模型在大规模日志分析、网络流量分类中取得显著成果[4],通过自动化特征提取显著提升了检测的准确性。然而,这些方法更多地用于离线分析,缺乏对实时性和持续对抗场景的有效支持,尤其是在动态防御和即时决策方面仍存在较大局限性。

1.3人工智能在网络安全中的应用

强化学习的环境互动和奖励机制表现出了较强的自适应能力,使得系统能够在面对未知攻击时快速响应,得以匹配上述动态、即时决策困难的现状[5]。其通过学习历史数据和实时交互构建高效的决策模型,可以在企业网络安全中广泛应用于动态威胁检测、攻击路径预测和自动化防御策略优化。

2 强化学习在网络攻防中的探索

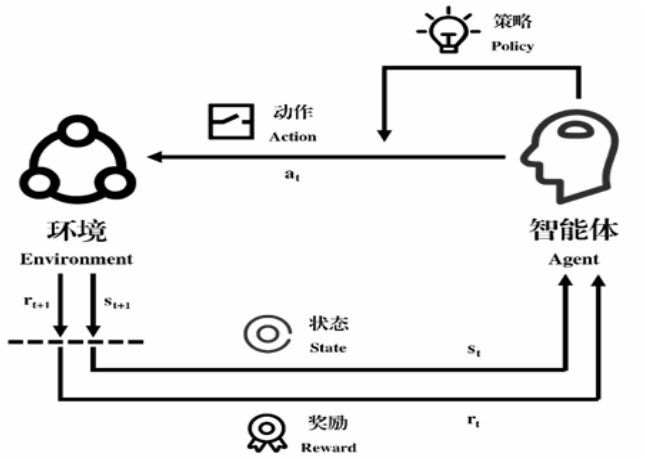

图1 关于强化学习的基本概念

2.1主流算法

目前,强化学习的主流算法主要分为两类:基于值函数逼近(Value Approximation-Based)和基于策略梯度(Policy-Based)的方法[6],分别适用于不同的情况,下面分别列举一些典型算法。

基于值函数逼近——深度Q网络(DQN)[7]

深度Q网络(Deep Q-Network,DQN)利用深度学习方法逼近Q值函数,是一种针对离散动作空间决策优化问题的强化学习算法。

深度Q网络通过近似Q值函数来更新状态-动作对的价值。DQN的核心公式包括Q值更新公式和目标函数,其中Q值更新公式如下：

$$Q(s_t,a_t)+\alpha\left[r_t+\gamma\max_{a'}Q(s_{t+1},a')-Q(s_t,a_t)\right]\rightarrow Q(s_t,a_t) \quad (1)$$

其中,$Q(s_t,a_t)$表示在状态 s_t 下选择动作的Q值；r_t 是在时间点 t 时获得的即时奖励；γ 是折扣因子,表示未来奖励的权重；$\max_{a'}$表示在下一状态s_{t+1} 下的最大Q值；α 是学习率,控制着学习过程中的步长。

假设有一个目标Q值网络和一个行为Q值网络,目标是最小化目标Q值和当前Q值之间的误差,DQN的目标函数如下：

$$L(\theta)=E(s_t,a_t,s_{t+1},a_{t+1})\left[\left(y_t-Q(s_t,a_t;\theta)\right)^2\right] \quad (2)$$

$$y_t=r_t+\gamma\max_{a'}Q(s_{(t+1)},a';\theta^-) \quad (3)$$

其中,θ 是行为Q值网络的参数；y_t 是目标Q值,θ^- 表示目标网络的参数。

基于策略梯度——近端策略优化(PPO)[8]

近端策略优化(Proximal Policy Optimization,PPO)是一种基于策略梯度的强化学习方法,其通过限制策略更新幅度(如引入KL散度或剪辑目标函数)实现了训练稳定性与效率的平衡。PPO的目标函数是：

$$L^{CLIP}(\theta)=E_t[\min(r_t(\theta)\hat{A}_t,clip(r_t(\theta),1-\epsilon,1+\epsilon)\hat{A}_t)] \quad (4)$$

其中,$r_t(\theta)$ 是策略比率,表示当前策略和旧策略在同一状态下采取同一动作的概率比；$\hat{A}_t$ 是优势函数(Advantage),表示某动作相较于平均水平的好坏程度；ϵ 是一个小的超参数,用于限制策略的变化。

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.

优势函数通常通过广义优势估计(Generalized Advantage Estimation, GAE)来计算：

$$\hat{A}_t = \sum_{l=0}^{\infty} (\gamma \lambda)^l \delta_{t+l} \quad (5)$$

其中，δ_{t+l} 是时刻 t 的时间差分(TD)残差；γ 是折扣因子，决定未来奖励的权重；λ 是GAE中的平衡参数，用于控制偏差与方差之间的权衡

PPO在动态防御策略的优化中表现出色，能够有效减少策略更新过程中的剧烈振荡，适应复杂或连续的动作空间场景。PPO在多阶段防御任务中展现了较好的性能，例如优化资源分配或动态调整安全策略，但其对高频交互任务的实时性支持仍需进一步提升。

综合方案——演员-评论家架构(A3C/A2C)[9]

演员-评论家架构(Actor-Critic)结合了策略优化与值函数估计，具备强大的稳定性和并行训练能力。其中的A3C(异步优势演员评论家)通过多线程并行显著加速了策略更新过程[9]，特别适合大规模复杂网络模拟环境中的迭代训练，其演员更新策略如下：

$$\theta_{actor} \leftarrow \theta_{actor} + \alpha_{\pi} \nabla_{\theta_{actor}} \log \pi_{\theta_{actor}}(a_t|s_t) \hat{A}_t \quad (6)$$

其中，$\pi_{\theta_{actor}}(a_t|s_t)$是当前策略在状态下 s_t 选择动作 a_t 的概率；$\hat{A}_t$是优势函数，表示当前动作相较于平均水平的好坏。

其评论家更新值函数如下：

$$\theta_{critic} \leftarrow \theta_{critic} + \alpha_V \nabla_{\theta_{critic}} \left(R_t - V_{\theta_{critic}}(s_t)\right)^2 \quad (7)$$

其中，R_t 是回报，通常通过蒙特卡洛方法或时间差分方法来估计；$V_{\theta_{critic}}(s_t)$则是评论家估计的状态值函数。

此外，A2C(同步版A3C)通过去除异步设计，进一步提高了训练效率，但在应对动态攻击路径时，仍需优化策略的实时响应能力。

2.2学界前沿方案

胡等人[10]考虑到了多攻击方群体的共谋攻击情况，通过结合随机博弈模型和A3C算法，并建立合作系数 μ 刻画攻击方合作的影响，利用多智能体表征群体性格特征。

Aditya等人提出了一种层次化多智能体强化学习方法，用于网络安全防御。首先定义每个智能体独立负责特定子网，使用共享奖励机制进行协作学习；再基于策略算法PPO将复杂的防御任务分解为子任务(如主机调查、恢复和流量控制)，由主策略选择合适的子策略，提升了协作效率。

经此启发可知，多智能体系统不仅能够有效建模攻击者之间的协同进攻行为，还能通过将庞大的网络拓扑划分为多个子网进行局部防御，从而为企业级大规模网络安全智能策略的实际应用提供了重要支持。

此外，注意到Cyber Operations Research Gym(CybORG)是一个良好且先进的网络防御拟真环境，非常适合用于强化学习驱动企业网络安全智能决策的方案设计和实验。

2.3工业界应用现状

在工业界，AI增强的网络安全运维系统已经在多个安全厂商和大型企业中得到了应用，但主流思路依旧是机器学习和深度学习。Vectra AI公司提到强化学习目前尚处于较新阶段，难以做到在生产情景下表现良好，该公司正试图开发强化学习智能体，用于生成供入侵检测系统(IDS)学习的数据。

2.4强化学习在企业网络安全场景落地的挑战

高维状态空间：企业实际场景中的网络状态、攻击行为及防御策略极度复杂多样，导致训练过程计算量大，计算资源消耗过多，遇到可扩展性问题；

样本效率低：在资源有限的网络安全环境中，收集足够的Label数据是较为困难的。一个典型是零日攻击，几乎没有机会找到历史数据，这将导致该场景的应用效果只能依赖模型的泛化能力，十分不稳定；

对抗性攻击：攻击者可能通过操纵训练数据来影响系统行为，进而导致次优的防御策略。

可解释性差和迁移性差：在企业网络安全中，决策透明性和从一个环境到另一个环境的知识迁移是提高系统效率和应对新型威胁的关键。

实时性需求：防御策略需要短时间内做出响应，对强化学习算法性能有较高要求。

克服这些挑战是强化学习在企业网络安全中广泛落地的关键。

3 强化学习的企业网络攻防模型

本研究从理论角度出发，提出了一种基于强化学习的企业网络攻防模型框架，旨在提升企业网络安全防御的智能化与动态响应能力。该框架融合了多智能体系统(Multi-Agent Systems, MAS)与分层决策机制(Hierarchical Decision-Making)，通过优化算法与知识增强策略，构建一个高效且可扩展的防御体系。

3.1多智能体架构设计

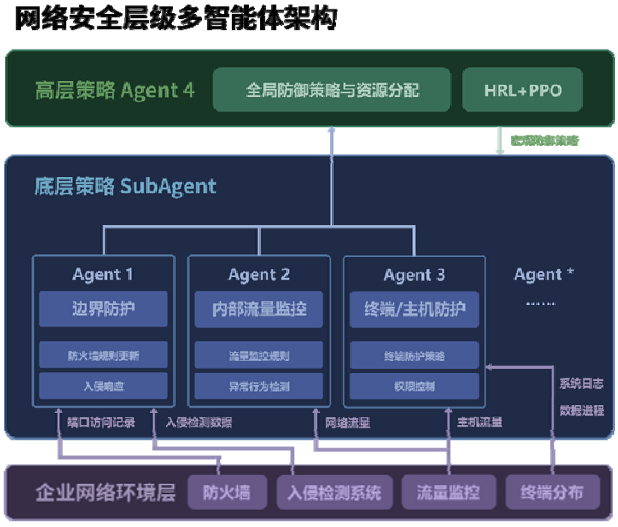

图2 网络安全层级多智能体架构

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.

本模型采用多智能体协作机制,部署在企业网络的不同区域或安全模块中,包括防火墙、入侵检测系统、流量监控和终端检测等。每个智能体(Agent)依据其职责分工,选用最适合的强化学习算法,以优化其防御策略。

3.1.1 Agent分工

具体而言,整个企业网络被划分为若干安全子模块,每个子模块由专门的强化学习智能体(Agent)负责管理与防护。例如:

Agent 1(边界防护)：专注于防火墙策略的动态调整和入侵检测系统规则的优化,实时监控外部入侵尝试并迅速响应。

Agent 2(内部流量监控)：负责监控内部网络流量,检测异常行为和可疑连接,防止横向移动和内部威胁的扩散。

Agent 3(终端/主机防护)：专注于终端设备的安全防护,包括主机隔离、进程监控和权限提升检测,确保单点防护的有效性。

Agent 4(高层策略Agent)：作为整体防御体系的统筹者,基于全局威胁态势进行资源分配和防御策略的宏观调整,协调各低层Agent的工作。

3.1.2 Agent算法选择

不同智能体在执行其特定任务时,适用的强化学习算法有所差异。根据智能体的职责和工作环境,建议采用以下算法:

Agent 1(边界防护)：由于边界防护需要处理离散动作空间,如调整防火墙规则和封禁IP地址,适合采用深度Q网络(DQN)。

Agent 2(内部流量监控)：内部流量监控涉及复杂且连续的流量数据分析,推荐使用近端策略优化(PPO)。PPO在处理连续动作空间和稳定性方面具有优势,能够有效识别和响应内部网络中的异常流量模式。

Agent 3(终端/主机防护)：终端防护需要高效处理多任务和多目标防御,适合采用演员-评论家(A3C)。A3C通过多线程并行训练,能够提升训练效率,适应多变的终端安全威胁。

Agent 4(高层策略Agent)：高层策略Agent需要综合多智能体的输出,制定全局防御策略,推荐使用分层强化学习(HRL)结合PPO。HRL能进一步分解复杂策略,提升全局决策的灵活性和精准性。

3.2分层强化学习机制

3.2.1高层(策略层)

高层策略Agent负责全局防御策略的制定与优化,其主要功能包括：

状态表示：整合网络全局态势信息,如关键节点的负载情况、告警分布(如IDS触发率)、网络拓扑的攻击面评估以及业务优先级信息。

动作空间：定义宏观防御策略动作,如选择防御模式(“严格模式”、“常规模式”、“限流模式”等)、在不同区域部署额外监控资源、对下级Agent分配权限或资源配额等。

奖励函数：基于全局攻击阻断率、误报率、资源占用和业务连续性等综合指标设计奖励函数,确保策略的优化方向符合整体安全目标。

3.2.2低层(执行层)

低层Agent负责具体防御动作的执行,其主要功能包括：

状态表示:聚焦于局部防护对象的信息,如某台主机的流量特征、系统日志、可疑进程或某段网络流量的统计特征。

动作空间：定义具体的防御动作,如封禁特定IP地址、隔离可疑主机、限制某端口的带宽使用、触发高强度身份验证等。

奖励函数：结合准确阻断攻击、避免过多误报、最小化资源消耗等因素,设计奖励机制,并确保与高层决策保持一致,形成分层奖励的递进结构。

3.3对抗性强化学习(Adversarial RL)

3.3.1模拟攻击者

在模型的训练阶段,引入模拟攻击者或对抗性Agent,通过模拟多种攻击手段(如DDoS、APT横向渗透、零日漏洞利用等),与防御Agent进行对抗训练。这种方式促使防御Agent在不断变化和进化的攻击环境中提升其策略的泛化能力和鲁棒性。

3.3.2训练方式

交替训练:在同一模拟环境中,进攻方与防御方轮流更新策略,确保双方在对抗中不断优化各自的策略。

自博弈训练:同时训练多个攻击与防御策略,通过并行训练形成多样化的对抗场景,丰富防御Agent的经验库,提升其应对复杂攻击的能力。

3.4知识增强(Knowledge-Enhanced RL)

知识图谱：构建企业安全知识图谱,整合网络拓扑、业务依赖关系、常见攻击路径和威胁情报等信息,增强Agent对网络环境和攻击模式的理解。通过将关键业务节点优先级、已知攻击指示器(Indicators of Compromise,IoC)和疑似受损链路等信息纳入状态表示,辅助Agent更快速、准确地定位和识别威胁。

特征降维与选择：应用主成分分析(Principal Component Analysis,PCA)或自动编码器(Autoencoder)等技术对高维状态空间进行降维,保留攻击判别度最高的特征,提升模型的训练效率和实时性。同时,将安全专家的经验规则或告警阈值嵌入状态特征,如“当前连接数远超正常均值”即视为警告标志,进一步优化状态表示。

3.5多维奖励机制

设计多维度的奖励机制,以全面评估防御策略的效果：

攻击阻断成功率(正向奖励)：成功检测并阻断攻击时给予正向激励。

误报代价(负向奖励)：误封合法用户或造成业务中断时给予负向激励。

资源消耗(负向奖励)：超过资源配额或占用过多网络带宽时设置惩罚项。

长期收益(综合奖励)：系统在较长时间内无重大安全事故时给予额外正向激励,引导Agent实现对整体安全最优的长期决策。

动态调整:根据企业的风险偏好或安全运营策略,动态调整不同奖励项的权重。例如,在高风险时期,可以适当放宽“误报

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.

率”要求,以优先阻断可能的威胁,确保防御策略的有效性和灵活性。

3.6实验模拟

CybORG提供了一个可扩展的仿真平台,支持网络攻防策略的测试与优化。它通过对抗性环境和多智能体协作,帮助研究者评估强化学习模型在真实世界复杂场景中的表现和鲁棒性。

4 结语

强化学习在企业网络防御中展示了巨大的潜力,特别是在自适应防御与恶意行为识别方面。传统的网络安全防御机制往往依赖于静态规则和阈值设置,难以应对日益复杂的网络攻击和不断变化的攻击手段。随着算力的不断提升和强化学习算法的逐步优化,强化学习则能够通过与环境的持续互动,自动优化防御策略,实时应对新的攻击模式,为企业提供持续的智能升级。

本文所提出的基于强化学习的企业网络攻防模型框架仍面临一系列挑战。首先,在大规模环境下的适用性存在问题。真实企业网络的设备数量庞大、网络结构复杂,流量规模和攻击事件频繁,远远超出实验室环境的模拟能力,这使得现有方案在实际应用中难以处理如此庞大的数据量和实时变化的网络环境；其次,实时性与资源消耗问题也不可忽视。强化学习模型通常需要大量计算资源进行训练和推理,尤其是在高并发情况下,可能会面临性能瓶颈,影响网络防御的实时响应；最后,可解释性不足是当前强化学习模型广泛应用的瓶颈之一。由于强化学习模型的决策过程通常为“黑箱”操作,安全运维人员难以理解和追踪其防御决策,这在实际生产环境中可能引发信任危机,导致实施困难。

未来,基于强化学习的企业网络防御模型将向更加智能和可行的方向发展。首先,多智能体与协作防御将成为趋势。通过在网络中不同节点或安全模块之间实现智能体的协同决策,能够更好地整合各个局部防护措施；同时,知识增强与可解释AI将成为强化学习发展的重要方向。将安全策略、专家经验、威胁情报等外部知识融入强化学习训练过程中,能够提升模型的决策能力、决策过程的透明度和可解释性,推动模型在实际生产环境中的落地和应用。

[参考文献]

[1]WANG Wenhao,SUN dingyuanhao,JIANG feng,et al.Research and challenges of reinforcement learning in cyber defense decision-making for intranet security[J].Algorithms,2022,15(4):134.

[2]张瑜,潘小明,LIUQingzhong.APT攻击与防御.清华大学学报(自然科学版),2017,57(11):1127-1133.

[3]张蕾,崔勇,刘静,等.机器学习在网络空间安全研究中的应用[J].计算机学报,2018,41(9):1943-1975.

[4]王伟.基于深度学习的网络流量分类及异常检测方法研究[D].合肥:中国科学技术大学,2018.

[5]冯瑞佳,荀洋,张考,等.基于人工智能的网络安全技术与应用[J].网络安全技术与应用,2024(3):118-121.

[6]刘全,翟建伟,章宗长,等.深度强化学习综述[J].计算机学报,2018,41(01):1-27.

[7]MNIHV,KAVUKCUOGLUK,SILVER D,et al.Human-level control through deep reinforcement learning.[J].Nature,2015,518(7540):529-533.

[8]张振,黄炎焱,张永亮,等.基于近端策略优化的作战实体博弈对抗算法[J].南京理工大学学报,2021,45(01):77-83.

[9]Volodymyr Mnih,Adrià Puigdomènech Badia,Mehdi Mirza,et al.Asynchronous methods for deep reinforcement learning[J].CoRR,2016.

[10]胡浩,赵昌军,刘璟,等.基于随机博弈与A3C深度强化学习的网络防御策略优选[J].指挥与控制学报,2024,10(1):47-58.

作者简介：

何俊(1983--),男,汉族,上海人,硕士研究生,工程师,主要研究方向为网络安全和算力网络。

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.