基于人工智能的网络威胁研究

Research on AI-Based Cyber Threats

陈　蓉,侯　捷,徐　雷(中国联通研究院,北京 100053)
Chen Rong,Hou Jie,Xu Lei(China Unicom Research Institute,Beijing 100053,China)

摘　要:
人工智能技术的快速发展为网络安全领域带来了革命性的变化。一方面,人工智能可以用于帮助发现网络威胁,提升安全防御能力。另一方面,人工智能也被用于增强攻击能力,加剧威胁风险。首先,梳理了人工智能的优势和在网络威胁发现工作中的研究进展。其次,总结了人工智能的内在特点和在攻防视角下的应用进展。最后,结合现有网络威胁领域的进展分析,对人工智能在网络威胁中的研究进行总结探讨。

关键词:
人工智能;网络威胁发现;人工智能安全
doi:10.12045/j.issn.1007-3043.2025.09.011
文章编号:1007-3043(2025)09-0058-05
中图分类号:TN915.08
文献标识码:A
开放科学(资源服务)标识码(OSID):

Abstract:
The rapid development of AI technology has brought revolutionary changes to the field of cyber security. On the one hand, AI can be used to help detect cyber threats and enhance security defense capabilities. On the other hand, AI is also used to enhance attack capabilities and increase threat risks. Firstly, the advantages of AI and its research progress in cyber threat detection are summarized. Secondly, the inherent characteristics of AI and its application progress from an offensive and defensive perspective are summarized. Finally, based on the analysis of the progress in the existing field of cyber threats, it summarizes and discusses the research on AI in cyber threats.

Keywords:
AI;Cyber threat discovery;AI security

引用格式:陈蓉,侯捷,徐雷. 基于人工智能的网络威胁研究[J]. 邮电设计技术,2025(9):58-62.

0 引言

人工智能技术的快速发展为网络安全领域带来了革命性的变化。一方面,人工智能技术可显著提升网络安全防御能力,例如通过机器学习算法实现对海量日志数据的实时分析,提高威胁检测的准确性与效率;利用深度学习模型自动识别软件漏洞,加强系统的事前防护能力;结合自然语言处理技术实现对安全事件的智能研判与自动化响应,大幅降低响应时间。另一方面,人工智能技术也正被攻击者广泛利用,例如通过生成对抗网络(GAN)制造高度逼真的虚假信息,发起智能社会工程攻击;利用强化学习构建自适应的恶意软件,可绕过传统检测机制;借助对话生成模型批量生成钓鱼邮件,显著提高攻击覆盖率和欺骗性。这种“双刃剑”特性使得人工智能在网络空间安全中的角色日趋复杂。因此,系统性地研究人工智能技术在网络威胁中的应用形态与相应的防御策略具有重大现实意义。本文围绕该主题展开了2方面的研

收稿日期:2025-07-11

究:一是总结人工智能在威胁发现中的最新研究进展,二是分析人工智能赋能的新型网络威胁模式与应对机制。

1 人工智能在网络威胁发现中的应用

1.1 人工智能的优势

传统的安全工具,比如入侵检测系统(IDS)和防火墙,长期依赖规则引擎这一核心机制。规则引擎通过匹配预定义的攻击签名或行为模式来识别威胁。然而,在当今的威胁环境下,这种模式的局限性愈发凸显[1]。

a)反应滞后。规则库的更新总是滞后于新威胁的出现,对于零日漏洞、变种恶意软件和高级持续性威胁(APT)的检测能力有限,安全响应始终处于被动状态。

b)易于规避。攻击者可通过代码混淆、加密通信或微调攻击手法,轻松绕过僵化的规则匹配,传统防御体系难以应对刻意规避行为。

c)告警疲劳。为追求覆盖率,规则往往设置得较为宽泛,导致产生海量误报,淹没真实的安全事件,极大消耗了安全分析师的精力,易造成关键信息遗漏。

d)维护成本高昂。规则库的构建、测试、发布与更新严重依赖专业安全人员的经验,需持续投入高技能人力进行手动维护,不仅运营成本巨大,也难以实现规模化敏捷响应。

人工智能,特别是机器学习与深度学习,为网络威胁发现带来了根本性变革。人工智能驱动的检测系统不再依赖僵化的规则,而是通过海量数据学习,构建能够识别“异常”的动态模型。相较传统规则,人工智能有以下几大优势[2]。

a)上下文感知与行为分析。人工智能能够融合网络流量、用户实体行为分析(UEBA)、终端日志与云环境数据,构建多维度“正常行为”基线,实现对偏离基线的微异常和高风险操作的实时检测。

b)持续学习与自适应。依托在线学习和增量训练机制,人工智能模型可不断从新型攻击样本、威胁情报和运营反馈中自动迭代优化,具备对抗环境演化与攻击手法变迁的自适应能力。

c)处理海量异构数据。人工智能能够以远超人工处理的速度和规模,对来自不同源头(如日志、流量、文件、身份信息等)的结构化与非结构化数据进行关联挖掘,从噪声中提取微弱攻击信号,实现早期威胁感知和横向移动追踪。

表1给出了传统规则的威胁发现与AI驱动的威胁发现的对比。

表1　传统规则的威胁发现与AI驱动的威胁发现对比

对比项	传统规则检测	AI驱动检测
检测原理	基于签名的静态模式匹配	基于数据的动态行为分析与异常检测
威胁覆盖	主要针对已知威胁,对未知、变种威胁效果不佳	有效识别已知与未知威胁,特别是零日攻击和APT
响应速度	依赖规则库更新,反应滞后	实时检测,快速响应
准确率	误报率高,易导致“告警疲劳”	误报率显著降低,告警质量更高
适应性	规则僵化,需手动维护	模型自我学习,动态适应新威胁
维护成本	规则库持续维护成本高	模型训练和调优成本高,但自动化程度高,长期运维成本可控

1.2 人工智能在威胁发现中的应用进展

人工智能技术在网络安全威胁检测中的应用主要体现在机器学习和深度学习2个方面。机器学习算法,如支持向量机(SVM)[3]、随机森林[4]等,能够从历史数据中学习正常和异常行为模式,用于检测已知和未知威胁。这些算法可以处理结构化数据,如网络流量、系统日志等,可通过特征工程提取关键信息,构建分类模型。在实际应用中,相较于传统规则学习,机器学习算法往往表现出较高的准确率和较低的误报率。在网络威胁发现中,机器学习主要用于异常检测、恶意代码检测、网络钓鱼检测、行为分析等[5]。

a)异常检测领域。机器学习模型能够高效分析包括网络流量、系统日志和用户行为模式在内的海量数据,从中识别偏离正常模式的异常行为并检测潜在威胁。通过对历史数据的学习,模型能够构建出对特定网络或系统中“正常”行为的精确理解,进而实现对异常活动或未知安全风险的智能标记与预警,大大提升了检测的覆盖面和响应速度。

b)恶意软件检测领域。机器学习方法不仅可以识别已知恶意软件,还能够通过对文件特征、代码结构及运行时行为进行深度分析,检测出新型、变种或经过混淆的恶意程序,有效弥补了传统依赖特征码的杀毒工具在应对零日威胁方面的不足。

c)网络钓鱼检测领域。机器学习模型能够综合解析电子邮件内容、URL特征、发信行为以及用户交互模式,通过自然语言处理和图关联分析等技术识别

伪装和欺骗企图,显著提高了对复杂网络钓鱼和鱼叉式钓鱼攻击的判断准确率。

d)行为分析领域。机器学习利用用户和实体行为分析(UEBA)系统,通过对用户、设备等实体行为的持续监控与建模,自动识别如数据异常访问、权限滥用、横向移动等内部威胁,可广泛适用于企业内部风险治理与账号泄露场景的及时发现与响应。

深度学习技术,特别是卷积神经网络(CNN)[6]和循环神经网络(RNN)[7],在处理非结构化数据和高维数据方面展现出强大优势。在网络安全领域,深度学习主要用于分析网络流量、检测恶意代码,以及与自然语言处理技术结合,用于分析文本数据,如检测恶意邮件、识别社交工程攻击等。在网络威胁发现中,深度学习主要用于恶意代码检测、入侵检测、漏洞利用检测以及威胁情报分析与预测等[9]。

深度学习技术在恶意代码检测中展现出显著优势。通过使用卷积神经网络(CNN)等模型,可以分析恶意代码的二进制特征和行为模式,从而识别出隐藏的恶意代码。深度学习算法能够自动提取高维度特征,超越传统基于签名的方法,检测未知和变种恶意软件。通过大规模数据训练,模型可以不断更新和优化,提高检测的准确率和鲁棒性,有效应对新型和复杂的恶意代码威胁。

深度学习技术通过使用循环神经网络(RNN)和长短期记忆网络[7](LSTM),可以分析网络流量和用户行为的时间序列数据,发现异常模式,在入侵检测系统(IDS)中得到广泛应用。深度学习模型能够识别复杂的攻击行为,如零日攻击和高级持续性威胁(APT),显著提升入侵检测的精度和响应速度。其自学习能力使得入侵检测系统可以适应不断变化的网络环境,提供更强的保护。

在漏洞利用检测中,深度学习技术通过使用自编码器和生成对抗网络[8](GAN)等模型,可以检测系统中的潜在漏洞和异常行为。深度学习算法能够自动提取和分析应用程序和网络流量的特征,识别利用漏洞的攻击模式。这不仅提高了检测的准确性,还能在攻击发生前进行预警,防止漏洞被恶意利用,减少系统被攻破的风险,保障网络安全。

深度学习技术在威胁情报分析与预测中也发挥着重要作用。通过使用自然语言处理(NLP)技术,深度学习模型可以从大量非结构化数据(如安全报告、社交媒体和暗网数据)中提取有价值的威胁情报。利用时间序列分析和预测模型,可以预测潜在的网络威胁和攻击趋势。深度学习的自适应能力使系统能够实时更新威胁情报,提高威胁识别的准确性和前瞻性,从而增强网络防御的主动性和预见性。

2 人工智能对网络威胁的赋能

2.1 人工智能的特点

人工智能对网络威胁的赋能作用取决于人工智能的内在特点。人工智能的内在特点可以增强恶意代码的能力,从而引发新的网络威胁。根据人工智能对人类能力的模拟,人工智能的能力可以分为感知、学习和认识3类[10]。感知是对外部刺激信息进行感知和加工,主要包括对文字、语音、图像、视频等连续或非连续信号的处理。学习表示在与环境或样例的交互中进行学习,主要包括无监督学习、有监督学习和强化学习等。认知代表从已有知识进行推理、规划和决策等,主要包括自然语言处理、知识表示等。当人工智能的感知、学习和认识等能力被应用到网络安全领域时,可以赋予网络威胁学习能力和决策能力[11]。学习能力指攻击者从结构化、非结构化数据中通过正负反馈学习数据的统计规律、特征分布或最优路径等,并将规律应用于攻击任务中。例如,人工智能可以从良性应用中学习其流量特征,并把恶意软件的流量特征转换为良性应用的流量特征,以规避基于流量的机器学习检测器。决策能力指攻击者可以根据已学习到的知识,对未知数据或事件进行处理、分析与决策,使恶意代码能够根据环境变化指导恶意活动的状态变化。例如,人工智能可以帮助恶意代码找到攻击目标,执行定向攻击;集群智能恶意代码可以根据环境的变化来确定节点状态和下一步的攻击任务等。

2.2 人工智能在网络威胁中的作用

人工智能在网络威胁中的作用主要包括伪造与欺骗、隐蔽与匿名、感知与决策、定向与定制以及规模化与自动化[11-12]。

伪造与欺骗是攻击方借助人工智能增强的学习能力,通过模仿其他系统的成员的静态特征和行为特征,来创造或改造恶意对象的相关属性,将恶意对象伪造成来自其他系统的成员,进而欺骗第三方检测系统或相关人员。代表性案例有通过改造恶意代码或流量的特征来绕过黑盒检测器、模仿他人邮件内容进行钓鱼以及伪造人像图像和音视频等。

隐蔽与匿名是攻击方借助人工智能增强的能力

及人工智能模型的性质,通过形态变换等方式,改造恶意对象的相关属性,将恶意对象混淆于良性应用中,以达到隐藏自身行为、增强溯源难度和规避检测的目的。隐蔽与匿名和伪造与欺骗有重合的部分。例如,通过学习良性应用的通信特征,改造自身通信特征的恶意软件,在实现伪造与欺骗的同时,也实现了自身的隐匿。其他案例有借助神经网络进行隐蔽寻址、借助神经网络模型隐蔽投递恶意载荷以及借助神经网络隐藏攻击意图等。

感知与决策是攻击方借助人工智能增强的学习和决策能力,对环境中出现的文字、图像、视频、音频及其他数字信号进行处理,以理解环境状态,并根据已学习到的知识和环境反馈,对下一步的行动给予指导。代表性案例有基于机器学习方法的验证码破解、防火墙规则探测、攻击目标选择以及集群智能僵尸网络等。

定向与定制是攻击方借助人工智能增强的能力及人工智能模型的性质,根据预设规则,有针对性地选择和识别攻击目标、定制开发攻击载荷,以达到攻击效果和收益的最大化。代表性案例有基于机器学习的鱼叉式钓鱼、定向网络攻击以及用户追踪等。

规模化与自动化是攻击方结合人工智能增强的能力,大规模、自动化地完成需要人工干预指导的攻击任务,进而扩大攻击影响范围,提升攻击收益。例如,自动化的鱼叉式钓鱼、防火墙规则探测以及凭证窃取等,都将需要人工干预的部分交给神经网络来完成,大大提升了攻击效率。当攻击者能力得到增强后,便能将这些能力应用于网络威胁中。

目前,随着大模型的加速发展,人工智能赋能攻击场景从而造成的安全威胁愈发明显。例如借助人工智能的钓鱼定位攻击,该攻击通过整合和关联网络空间中大量的个人信息[13],尤其是社交网络平台,对目标实施钓鱼攻击[14],进行定位和追踪,进而采取具有针对性的攻击方法。或者攻击者通过生成虚假的信息,诱使目标进入预先设定的陷阱中,使目标完成攻击者设定的操作。借助人工智能在学习文字[15]、图像[16]、语音[17]和视频[18]等领域的能力,攻击者能够更快地生成更具有欺骗性的信息。在软硬件层面,人工智能可以实施防火墙绕过。Web应用防火墙(Web application firewall,WAF)通常部署在系统边界,能够拦截恶意带有攻击载荷的非法访问数据,保护应用系统的安全。近些年,有研究人员使用人工智能技术自动化地探测WAF规则,生成可以绕过防火墙的攻击载荷[19]。人工智能能助力漏洞发现与利用的过程。目前已有研究人员通过人工智能帮助寻找和利用系统或程序漏洞,使用包括0 day在内的各种方式触发目标程序漏洞,获取目标系统权限[20]。

2.3 人工智能赋能网络威胁的防御进展

针对当前网络空间面临的一些威胁,安全人员采取了一系列防御措施。目前的防御措施主要分为两大类,分别是被动防御和主动防御技术。被动防御包括防火墙、入侵检测、恶意代码扫描和网络监控等技术,可以及时发现威胁和脆弱点,降低系统安全风险。主动防御包括零信任、数据加密、蜜罐、审计和态势感知等技术,可以对整个信息系统进行实时监控,捕捉网络及系统异常,对可疑行为进行告警,提升信息系统面对安全威胁时的响应速度。由于攻击者要从目标系统窃取情报或者对目标系统进行破坏、利用目标系统进行恶意代码传播等,要经过信息搜集、漏洞发现、漏洞利用等一系列步骤,如果信息系统能做好日常安全防护,降低安全风险,也能抵御一些已知的攻击。

此外,需认识到人工智能赋能下的网络威胁防御离不开“人”的作用。目前一些AI赋能的特定类型的网络攻击仍是网络安全的一个重要威胁,如口令探测和防火墙绕过。但对于这类攻击,人工智能做的是把已有的攻击工作自动化和高效化,主要作用是提升了攻击的效率,它对攻击的危害和防御本身并没有本质影响,已有防御手段仍能对人工智能赋能的网络威胁进行有效的防御。换句话说,从技术角度讲,防御仍是有效的。而不能防御住人工智能赋能的网络威胁的场景通常和人密切相关,如目标定位涉及到人在各类网络平台发布的各种信息,敏感信息恢复取决于人将信息破坏到什么程度,信息伪造更是针对人而进行的社会工程攻击。由此可见,人是网络攻击中的薄弱一环。有人参与的环节往往会具备独特的不确定性和脆弱性。“安全是三分技术七分管理”,减少网络安全中的包括人为因素在内的不确定性也是防御工作的一个方向。

3 总结与展望

人工智能在网络威胁发现中的应用极大地提升了网络安全防御的效率与准确性,其强大的数据处理能力和模式识别能力使其能够从海量网络数据中快

速识别出潜在的威胁行为,为安全分析师提供了有力的决策支持。然而,人工智能技术的双刃剑特性使其在网络威胁领域的应用同样带来了新的挑战,人工智能的脆弱性意味着同样可以将人工智能引入网络威胁中,将网络威胁智能化。攻击者可以利用人工智能生成更加复杂和隐蔽的攻击手段,如自动化攻击脚本、智能钓鱼邮件和深度伪造内容等,这些攻击不仅难以被传统防御手段所识别,还可能绕过基于人工智能的检测系统。目前,基于人工智能的网络威胁场景将随着新技术的出现不断更新。一方面,人工智能与传统网络安全技术结合,将进一步加剧现有网络安全的风险,甚至将网络威胁带到物理世界。另一方面,新兴技术和概念不断出现,使人工智能与新兴技术的结合可能产生新的网络安全威胁。因此,未来在人工智能与网络威胁的研究中,需要进一步优化和提升人工智能在威胁发现中的性能,如通过改进算法模型、增强数据质量和提升模型鲁棒性等手段,使其能够更好地适应不断变化的网络威胁环境;另外,也需要加强对人工智能赋能的网络威胁的防御研究,探索新的防御策略和技术,如对抗性机器学习、可解释人工智能和人机协同防御机制等,以有效应对人工智能带来的新型网络威胁。

需要认识到,人工智能本身是一个数据分析方法,是一个能兼容多种数据类型、发现数据规律的分析方法,而不是解决所有问题的“良药”,未来攻防博弈双方仍将互相促进,互相牵制,互相发展,但人工智能仍是攻防博弈中不可或缺的工具。期待未来能够构建更安全、更可靠的人工智能,助力网络安全不断健康发展。

参考文献:

[1] HUSAIN S, DEHGHANTANHA A. AI-driven security operations centers: a survey and a vision for the future[J]. IEEE access, 2021, 9: 87654-87671.

[2] XU H X, WANG S N, LI N K, et al. Large language models in cybersecurity: a comprehensive survey and evaluation[EB/OL]. [2025-02-02]. https://arxiv.org/abs/2405.04760.

[3] CHEN P H, LIN C J, SCHÖLKOPF B. A tutorial on ν-support vector machines[J]. Applied Stochastic Models in Business and Industry, 2005, 21(2): 111-136.

[4] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[5] 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.

[6] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[7] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[8] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[9] 刘宝旭,李昊,孙钰杰,等. 智能化漏洞挖掘与网络空间威胁发现综述[J]. 信息安全研究,2023,9(10):932-939.

[10] 邱锡鹏. 神经网络与深度学习[M]. 北京:机械工业出版社,2020.

[11] 王志,尹捷,崔翔,等. 人工智能赋能网络威胁研究进展[EB/OL]. [2025-02-02]. https://jcs.iie.ac.cn/xxaqxb/ch/reader/view_abstract.aspx?flag=2&file_no=202204120000001&journal_id=xxaqxb#.

[12] CEARLEY D, JONES N, SMITH D, et al. Top 10 strategic technology trends for 2020[EB/OL]. [2025-02-02]. https://iatranshumanisme.com/wp-content/uploads/2019/11/432920-top-10-strategic-technology-trends-for-2020.pdf.

[13] 刘东,吴泉源,韩伟红,等. 基于用户名特征的用户身份同一性判定方法[J]. 计算机学报,2015,38(10):2028-2040.

[14] BAHNSEN A C, TORROLEDO I, CAMACHO L D, et al. DeepPhish: simulating malicious AI[EB/OL]. [2025-02-02]. https://albahnsen.wordpress.com/wp-content/uploads/2018/05/deepphish-simulating-malicious-ai_submitted.pdf.

[15] AI安全小学生11号. 学好AI后,我手写技能直接废了[EB/OL]. [2025-02-02]. https://mp.weixin.qq.com/s/Og7hshjr7KR2oOzMCqg-Cw. 2021.

[16] YU C, BIN M, ZHUO M. Biometric authentication under threat: liveness detection hacking[EB/OL]. [2025-02-02]. https://i.blackhat.com/USA-19/Wednesday/us-19-Chen-Biometric-Authentication-Under-Threat-Liveness-Detection-Hacking-wp.pdf.

[17] REBRYK Y, BELIAEV S. ConVoice: real-time zero-shot voice style transfer with convolutional network[EB/OL].[2025-02-02]. https://arxiv.org/abs/2005.07815.

[18] Twitter (Mikael Thalen)[Z/OL]. [2025-02-02]. https://twitter.com/MikaelThalen/status/.

[19] Black Hat via YouTube. AutoSpear: towards automatically bypassing and inspecting Web application firewalls[EB/OL]. [2025-02-02]. https://www.classcentral.com/course/youtube-autospear-towards-automatically-bypassing-and-inspecting-Web-application-firewwalls-184943.

[20] Isao Takaesu. Deep exploit[EB/OL]. [2025-02-02]. https://github.com/13o-bbrbbq/machine_learning_security/tree/master/DeepExploit. 2018.

作者简介:

陈蓉,工程师,硕士,主要从事网络安全规划工作;侯捷,工程师,硕士,主要从事网络安全规划工作;徐雷,正高级工程师,博士,主要从事网络安全规划与指导工作。