人工智能视域下网络数据安全风险与防护策略研究

李晓娟，张敏
陕西省网络与信息安全测评中心，陕西 西安 710065
DOI: 10.61369/TACS.2026090009

摘　　要： 生成式人工智能与各类智能算法广泛应用，重构了网络数据采集、存储、流转、分析全流程。人工智能有效提升数据治理、安全检测和风险研判能力，推动传统被动防护转向主动预判式智能防御。与此同时，算法黑箱、数据投毒、对抗攻击、提示词注入等新型威胁持续涌现。AI 具有双面属性，在赋能防护的同时，也被攻击者用于自动化渗透、批量窃取伪造数据，显著增加数据泄露、篡改、滥用风险。本文结合 AI 发展现状，梳理其在网络数据安全领域的应用价值，归纳智能化场景下各类新型安全隐患，从技术、管理、制度维度剖析风险根源，从技术防护、制度建设、风险治理、人才培养、监管协同五方面构建适配 AI 时代的防护体系，为网络数据安全治理与智能防御建设提供理论和实践支撑。

关 键 词： 人工智能；网络数据安全；智能防御；数据风险；安全治理

Research on Network Data Security Risks and Protection Strategies from the Perspective of Artificial Intelligence

Li Xiaojuan, Zhang Min
Shaanxi Province Network and Information Security Evaluation Center, Xi'an, Shaanxi 710065

Abstract: The widespread application of generative artificial intelligence and various intelligent algorithms has reconstructed the whole process of network data collection, storage and circulation. Artificial intelligence effectively improves capabilities in data governance, security detection and risk assessment, shifting traditional passive protection to proactive and predictive intelligent defense. At the same time, new threats such as algorithm black boxes, data poisoning, adversarial attacks and prompt injection keep emerging. AI is double-edged: while empowering defense, it is exploited by attackers for automated infiltration and mass data theft and forgery, greatly raising risks of data leakage, tampering and abuse. Combined with the development status of AI, this paper sorts out its value in network data security and summarizes new hidden dangers in intelligent scenarios. It analyzes risk sources from technical, management and institutional dimensions, and builds an AI-adapted protection system covering technical defense,institutional improvement,risk governance, talent training and coordinated supervision, offering theoretical and practical support for network data security governance and intelligent defense construction.

Keywords: artificial intelligence; network data security; intelligent defense; data risk; security governance

一、绪论

（一）研究背景

数字经济背景下，网络数据已成为社会治理与产业发展的核心生产要素。大数据、云计算与人工智能的深度融合，使网络数据呈现海量化、多模态、高速流转的特征，传统静态网络安全防护模式面临迭代压力[1]。传统安全防护依赖固定特征库与人工研判，存在误报漏报率高、响应滞后、新型攻击识别能力弱、运维成本高等短板[2]。人工智能依托自主学习与行为关联分析优势，可实现网络异常行为、入侵攻击、违规数据流转的自动化识别与预警，有效弥补传统防护体系的局限性[3]。

与此同时，AI 技术的普及催生了自动化、智能化新型网络攻击，显著提升攻防对抗复杂度。《网络数据安全风险评估办法》的正式实施，明确了数据分层评估、常态化自查、专项动态评估的合规要求，构建了法治化、标准化的数据安全监管体系[4]。新规对 AI 数据处理、智能安全运维、风险闭环整改提出严格约束，在此背景下，开展人工智能场景下网络数据安全风险与防护策略研究，具有重要的现实意义。

（二）研究意义

1. 理论意义

目前整个学术界对网络安全和数据安全已经形成比较成熟的理论体系，但在人工智能引发的新型攻防对抗、算法层面的安全隐患及智能化场景下的特有风险方面，还有待于进入系统性深入研究[5]。现有成果多侧重于 AI 作为单一工具带来的增益效果，或是孤立地讨论 AI 的潜在危害，较少的从技术本身的双刃剑属性出发进行整体研究。本文尝试分析 AI 在赋能安全防护与诱发新型风

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

险之间的内在关联，在此基础上构建符合智能时代需求的数据安全治理框架，推动人工智能与网络安全交叉领域的理论发展，填补现有研究在系统性整合方面的空白[6]。

2. 实践意义

本文结合最新数据安全合规要求，系统归纳 AI 衍生的新型数据安全风险，从多维度提出可落地的防护方案。研究成果可指导政企单位规范 AI 数据处理流程、搭建智能防御体系、抵御智能化网络攻击，有效规避数据泄露与合规风险，为行业智能数据安全治理提供实操参考[7]。

（三）国内外研究现状

国外 AI 与网络安全融合研究起步较早，NIST、ISO 等机构建立了完善的智能安全评估标准，研究重点集中于机器学习异常检测、对抗攻击防护、智能入侵识别等领域，形成了成熟的技术研究体系。国内研究聚焦大模型在漏洞挖掘、风险预警、安全研判中的落地应用，随着数据安全法规持续完善，研究方向逐步从单一技术防护转向合规综合治理。但现有研究整体滞后于技术迭代，对大模型隐蔽风险、AI 合规盲区、新型智能攻击的研究不够系统，防护体系与法定评估要求适配性不足。

（四）研究内容与创新点

1. 研究内容

本文梳理人工智能在网络数据安全领域的核心应用价值，归纳智能场景下的新型数据安全风险，从技术、数据、管理、人才维度剖析风险成因，构建适配 AI 时代的多维一体化数据安全防护体系与落地路径。

2. 创新点

目前学界对传统网络安全与数据安全已有较完整的研究基础，但是现有成果大多从单一视角切入，要么关注 AI 的技术赋能效应，要么聚焦其潜在危害，很少将技术的两面性纳入统一框架加以审视。本文立足人工智能“既能加固防线、也能助长攻击”的双重特征，着重剖析大模型与智能算法衍生的隐蔽型新风险，推动智能防护能力与合规监管要求的有效衔接。

二、人工智能在网络数据安全领域的应用价值

（一）智能威胁检测，提升风险识别精度

人工智能通过引入深度学习与流量分析手段，建立覆盖网络访问、数据传输、系统操作的动态行为基线，能够自主发掘异常特征，精准识别内网渗透、恶意爬虫、越权访问、隐蔽隧道传输等各类风险，有效降低监测过程中的误报与漏报比例。

（二）自动化安全响应，缩短攻防响应时差

网络攻击突发性强、迭代速度极快，智能安全平台能够实现风险告警、流量清洗、攻击拦截、漏洞扫描、违规阻断等环节的全流程自动化响应，将攻防处置时间压缩至秒级，有效克服传统防护模式下响应迟缓、处置效率不高的弊端，大幅增强系统的动态防御能力。

（三）智能化数据治理，规范数据流转安全

人工智能可以自动执行全网数据的分类分级、脱敏处理和权限管控，准确划分公开信息、常规业务数据与核心敏感数据，覆盖数据从采集、传输、共享到销毁的整个生命周期。通过实时识别违规流转和越权操作行为，从源头阻断数据泄露、非法共享与隐私滥用等问题，切实提升数据治理的规范化程度。

（四）智能漏洞挖掘，实现主动预判防御

借助大模型的代码审查与智能推演功能，人工智能可以自动发现系统代码缺陷、配置漏洞及固件安全风险，提前预判业务系统与数据平台的潜在隐患。这一模式促使安全工作从传统的事后处置，转向事前预判、事中管控、事后复盘相结合的全周期主动防御。

三、人工智能时代网络数据安全新型风险

（一）训练数据缺陷风险

AI 模型的防御能力很大程度上取决于训练数据的质量，数据好坏直接决定了智能防护的效果[8]。如果训练数据存在噪声、偏差、过时残缺或来源不合规等问题，就容易导致模型判断出错、攻击识别失效。此外，许多机构的训练数据尚未完成全面脱敏，如果管控的不到位，可能引发大规模数据泄露。

（二）智能化网络攻击风险

人工智能大幅拉低了网络攻击的技术门槛，攻击者利用开源的 AI 工具，可以实现自动化批量扫描、智能破解、精准钓鱼和数据伪造。相对于传统攻击手段，AI 攻击隐蔽性更强、迭代更快、规模更大、溯源更难，能够轻松绕过传统安全设备的防护，窃取或篡改核心网络数据，严重冲击数据的保密性、完整性与可用性。

（三）算法与模型自身安全风险

大模型普遍存在算法黑箱问题，内部决策逻辑难以解释，如果发生误判或漏判，很难迅速溯源并加以修正。对抗样本攻击和提示词注入攻击可以绕过现有的安全校验，诱导模型输出涉密信息或执行非授权操作。当前针对模型漏洞的检测标准还不够完善，检测能力也存在不足。

（四）合规与管理体系滞后风险

现有安全管理制度更多适配传统网络环境，缺少针对 AI 数据处理与大模型应用的专项合规体系，难以对接《网络数据安全风险评估办法》的法定要求。行业内普遍存在 AI 权限过大、人机责任边界模糊、运维审计缺失、整改闭环不到位等问题。智能攻击链路日益复杂，取证困难重重，传统监管手段无法实现动态管控和有效追责，合规风险持续累积。

四、网络数据安全风险成因分析

（一）技术层面：智能防御迭代滞后

当前，人工智能领域的攻防博弈正陷入一种非对称困境。攻击技术的进化周期已大幅压缩，而既有防御机制的迭代却显迟滞。现有的智能安全模型，大多因对抗训练样本匮乏、特征提取维度受限，导致其泛化能力堪忧，面对数据投毒、对抗样本欺骗

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

及提示词注入等隐蔽威胁时，表现出较高的漏报率。

（二）数据层面：数据治理体系不完善

网络数据来源繁杂、流转频繁，多数单位未建立完善的数据分级分类、脱敏隔离、权限管控机制，训练数据与业务数据边界模糊。数据采集无授权、存储不加密、共享无审核、销毁不彻底等问题普遍存在，为数据泄露、数据投毒、数据滥用提供可乘之机。

（三）管理层面：AI 安全制度缺失

多数单位引入 AI 系统后，未建立配套的算法准入、模型测试、接口审计、上线运维管理制度，人机安全权责划分不清，缺乏大模型提示词、输出内容的常态化审查机制，安全运营无法覆盖 AI 全生命周期。

（四）人才层面：复合型安全人才短缺

AI 网络安全属于交叉学科领域，要求从业人员兼具网络攻防、数据治理、算法原理与合规管理能力。当前行业人才断层明显，传统运维人员缺乏 AI 风险识别能力，技术研发人员不熟悉安全攻防与合规体系，导致智能安全建设与风险处置能力整体不足。

五、人工智能视域下数据安全防护优化策略

（一）构建动态智能防御技术体系

搭建全域智能防御平台，整合流量分析、行为建模、漏洞挖掘、智能预警模块，实现风险实时感知、自动拦截、闭环处置。常态化开展模型对抗训练，提升模型对对抗攻击、数据投毒、提示词注入的识别精度。建立模型动态迭代机制，持续适配智能攻击迭代趋势，实现从被动拦截向主动预判防御转型。需要落实 AI 模型全生命周期管控，强化接口访问控制、输入输出过滤，采用联邦学习、差分隐私等技术降低数据泄露风险，实现操作可审计、事件可溯源。

（二）健全数据全生命周期安全治理

紧盯数据全生命周期管控，把分类分级落到实处，对一般、重要及核心敏感数据逐一明确防护策略，不搞“上下一般粗”。严把 AI 训练“入口关”，坚持“非必要不授权”，从严限定模型调取数据的范围和权限，把共享、导出、销毁各环节的审核链条串起来，切实堵住源头泄密漏洞。

（三）强化合规监管，压实主体责任

以《网络数据安全风险评估办法》为基准，当务之急是细化 AI 场景的合规颗粒度，厘清智能数据处理与模型应用的权责边界。机制设计上需分层施策：重要数据相关业务锁定年度强制评估，遇重大业务变更则触发专项排查，一般数据依托常态化自查实现风险前置。

（四）完善内部安全管理制度体系

着眼 AI 系统全生命周期管控，加快织密制度笼子，构建算法准入审查、模型上线审批及常态化监测的刚性约束。重点厘清人机交互场景下的安全责任边界，完善提示词管控、输出内容审核及全链路操作审计机制。在此基础上，推行风险自查常态化，实施台账清单化管理。

（五）培育复合型智能安全人才队伍

着眼学科交叉融合，推动高校增设网安与 AI 的复合型专业方向，以订单式培养破解人才供需错位难题。压实企业主体责任，将智能安全技能培训嵌入日常运维体系，着重提升一线人员对新型 AI 风险的实战化应对能力。同时，构建开放共享的行业交流机制，促进攻防案例与威胁情报的有序流动，深化政企联动及与测评机构的技术协作，以此撬动行业整体防护能力的系统性提升。

六、结论与展望

人工智能铺开以后，网络数据安全的防护路子彻底变了。它把过去那种“预判跟不上、响应慢半拍”的老毛病给治了一部分，推着治理往智能研判、动态调优的方向走[9]。但数据投毒、对抗攻击、提示词注入等，给现下的数据安全治理添了不少挠头的新难题。

在 AI 深度赋能与监管日趋严格的背景下，网络运营主体需摒弃单一技术防护思维，严格对标现行数据安全法规，搭建“智能技术防御 + 全周期数据治理 + 规范化制度管控 + 精准化监管约束 + 复合型人才支撑”的一体化防护体系，平衡技术创新与安全可控[10]。未来多模态大模型与智能体技术将持续迭代，数据安全风险将更加多元、隐蔽。后续需持续优化智能防御算法、完善合规治理体系、强化跨领域协同防护，持续筑牢 AI 时代网络数据安全屏障。

参考文献

[1] 陈明旭，周思远 . 大模型赋能网络安全智能防御机制研究 [J]. 计算机应用研究 ,2024,41(07):2089-2094.

[2] 郭星辰 .AI 智能化网络攻击的特征与溯源难点分析 [J]. 信息安全与通信保密 ,2024(05):67-73.

[3] 彭睿杰 . 人工智能算法安全风险研究综述 [J]. 计算机工程与应用 ,2023,59(18):36-44.

[4] 国家互联网信息办公室 . 网络数据安全风险评估办法 [Z].2026.

[5] 赵文博 . 智能时代网络数据安全治理体系构建研究 [J]. 数字治理研究 ,2024(03):78-85.

[6] 蔡明宇 . 我国智能数据安全治理的短板与优化策略 [J]. 法治与信息化 ,2024(02):33-39.

[7] 沈泽宇 . 人工智能赋能网络攻击的技术机理与风险防控 [J]. 信息网络安全 ,2024(03):44-51.

[8] 王铭远，陈思远 . 生成式人工智能的数据安全风险及治理路径 [J]. 信息安全研究 ,2024,10(02):156-162.

[9] 韦俊辰 . 生成式人工智能 Prompt 注入攻击风险及防护 [J]. 网络安全技术 ,2024,12(02):77-82.

[10] 林立峰 . 企业智能化数据安全防护体系建设实践 [J]. 网络安全与数据治理 ,2024,02(03):56-61.

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License