智能体驱动的网络安全和数据安全防护路径探究

唐伟
四川省建筑设计研究院有限公司，四川 成都 610000
DOI: 10.61369/TACS.2026020007

摘　　要： 在网络攻击越来越智能化、隐蔽化的趋势下，传统的安全防护模式已经不能适应复杂多变的安全环境了，智能体是摆脱防护困境的一种重要方式。本文主要研究智能体驱动的安全网络、数据安全防护，对目前的防护过程中存在的场景适配问题、决策可信度低、协同效率差、自身安全性差等问题进行分析，并从场景建模优化、可解释框架建立、协同架构搭建、内生安全加强四个方面提出防护方案，最后通过真实的项目案例来检验路径是否可行。对相关的方法论进行了系统的论述，给智能体在网络安全和数据安全领域规范化的使用提供理论基础和实践指导，提高防护的智能化程度，适应数字化转型过程中安全、合规的要求。

关 键 词： 智能体驱动；网络安全；数据安全；防护路径

Exploration on Agent-Driven Protection Paths for Cybersecurity and Data Security

Tang Wei
Sichuan Institute of Architectural Design Co., Ltd., Chengdu, Sichuan 610000

Abstract: Against the trend of increasingly intelligent and covert cyberattacks, traditional security protection models can no longer adapt to the complex and volatile security environment. Agents have become an important approach to addressing protection dilemmas. This paper focuses on agent-driven network security and data security protection, analyzes existing problems in current protection such as poor scenario adaptation, low decision-making credibility, weak collaboration efficiency, and insufficient self-security. It proposes protection schemes from four aspects: optimized scenario modeling, construction of interpretable frameworks, design of collaborative architectures, and enhancement of endogenous security. Finally, the feasibility of the paths is verified through real engineering projects. This paper systematically discusses relevant methodologies, provides theoretical basis and practical guidance for the standardized application of agents in cybersecurity and data security, improves the intelligence level of protection, and meets the requirements of security and compliance in digital transformation.

Keywords: agent-driven; cybersecurity; data security; protection paths

随着数字技术和全球化的不断推进，云计算、物联网等技术的发展使得网络攻击面越来越大，生成式人工智能（GAI）驱动的新型威胁频繁出现，数据泄露、勒索软件等安全事件时有发生，传统的被动防御方式已经不能满足复杂的、动态变化的安全形势需要了。与此同时，由于《中华人民共和国数据安全法》等法规不断更新，安全防护标准越来越高，但是安全运营效率低、误报率高以及依靠专家经验的问题仍然存在。在此背景下，智能体依靠自身的感知和推理决策的能力来解决安全防护的问题，它在安全场景中已经取得明显的效果。因此研究以智能体为驱动力的网络安全、数据安全防护路径，对解决目前存在的防护缺陷问题具有重大的理论和现实意义[1]。

一、智能体驱动的网络安全和数据安全防护路径探索难点

（一）场景适配不足 泛化能力受限

智能体在网络安全和数据安全场景下普遍存在着适配性弱的问题，现有的模型大多是以通用的环境进行训练的，并不能很好地理解异构网络结构、多种的数据格式以及不同的业务流程。复杂的攻防对抗环境中，智能体很难迅速应对攻击方式的更新以及业务环境的变化，因此它的决策鲁棒性和泛化能力不能达到实际应用的要求[2]。模型对于未知威胁、变种攻击的识别依靠高质量样本来支撑，在样本少、标签缺乏的边缘安全环境下容易造成误判或者漏判，自主学习和动态优化机制不够完善，使得智能体在实际应用中稳定性差，不能长时间地承担高强度、高动态的安全防护工作。

（二）决策逻辑模糊 可信程度偏低

智能体大多依靠深度学习来推动自己的发展，在决策的过程中是黑箱的，对于安全决策依据、推理路线以及风险评判逻辑等事项很难加以追踪。网络攻击阻断、数据访问控制、异常行为处置等重要环节中不能给出可以被解释、可以被核实、可以被审计的决策依据，从而降低安全运营人员对于防护结果的信任程度。黑箱决策和安全合规的要求相矛盾，不能达到责任划分、事件追踪、风险复盘的目的。决策结果并没有统一的标准来衡量它的优劣，由于误操作所造成的业务中断、数据异常访问等次生风险会严重阻碍智能体在高安全环境下的大规模应用[3]。

（三）协同机制缺失 联动效率低下

多智能体之间没有形成标准的协同架构和统一的调度方式，在跨域、跨层、跨系统的安全防护上存在着信息壁垒。各个智能体在威胁情报共享、态势感知汇聚、应急响应联动上很难形成高效的合力，会存在感知滞后、决策冲突、执行脱节等状况。分布式部署环境中由于节点之间通信延迟大、数据格式不同、权限不能协调等造成的状况十分明显，分布式部署环境不能完成整个网络安全态势的即时共享和协同防御。单个智能体的能力有限，群体智能的优势不能充分发挥出来，在大规模、分布式、协同式的攻击面前，整个防护体系的反应速度慢，处理的效果也差强人意[4]。

（四）安全漏洞凸显 自身风险突出

智能体在模型训练、数据交互、指令执行等环节存在着固有的安全脆弱性，很容易被当作新的攻击对象。对抗样本攻击、数据投毒、模型窃取等方式可以对智能体的决策逻辑造成干扰和操纵，从而使得防护失效或者反向控制。智能体的运行依靠大量的业务数据和敏感信息，数据采集、传输、存储过程中泄漏的风险以及篡改的风险明显增大[5]。自身安全防护体系不健全，缺少完善的权限控制、行为审计和异常隔离措施，一旦出现漏洞就会造成整个安全防护体系的崩溃，从而产生系统的安全威胁。

二、智能体驱动的网络安全和数据安全防护路径设计方式

（一）深耕场景建模 优化泛化能力

根据网络安全和数据安全业务实际运行规律来创建多维场景模型，剖析异构网络架构、动态业务流程以及多层数据权限体系的结构化特征并加以抽象，从而构建起一种态势表征方式。对威胁演化规律、数据流动路径、异常行为模式进行深入挖掘，创建符合实际防护需要的知识体系和规则约束，给智能体赋予稳定的决策支撑[6]。改善智能体的感知、推理和执行结构，提高模型应对动态对抗环境时的自适应调节水平，削减对标注样本过多地依靠。

以“试用部署的深信服安全产品体”项目为例，工作人员着重运用深信服安全智能体，与公司正在使用的深信服网络防火墙（NGAF）、全网行为管理（AC）、企业杀毒 EDR 等设备进行联动，结合勘察设计企业行业特有的异构网络架构、多层次数据权限体系以及业务动态流转特点，全面开展全场景建模工作。工作人员首先梳理集团总部和二级单位的网络拓扑、业务流程以及数据流动路径，整合各深信服产品的防护能力，提炼主要场景的安全特性，构建适配的防护知识体系与规则限制。同时，以深信服安全智能体为核心，改进其感知和推理结构，依托深信服安全大模型，整合各设备演练日志等信息进行专项训练，借鉴“云地协同、持续进化”框架创建在线学习体系，实现智能体与各产品的深度联动，使演练防护贴合日常运维，更具实操针对性。

（二）构建可解释框架 提升决策可信

把安全领域的知识、合规的要求同模型的推理机制结合起来，创建起可以被解释、追踪并加以验证的智能体决策系统，明晰风险评判、行为剖析以及处置命令发出时所涉及的逻辑链条，让防护行动具有明显的表达形式。建立决策依据的记录、留存和审计制度，保证安全事件全过程有迹可循、有据可查，满足监管审查、责任认定以及事件复盘等各方面的要求。利用结构化的规则和量化的指标来约束模型的输出，减少黑箱模型所造成的不确定因素，从而提高安全运营主体对于智能体决策结果的信任度[7]。

以“某银行 AI 反欺诈智能防护”为例，工作人员要将金融安全领域的知识、监管合规的要求同 AI 智能体的推理机制融合起来，创建起可以被解释、追踪并能核查的智能体决策架构。工作人员需要对银行交易场景中风险判定标准、异常行为特征和合规审查要求进行梳理，把它们转换成结构化的规则，嵌入智能体的决策流程当中，明确智能体从交易数据采集、风险特征分析、欺诈判定、处置指令生成整个逻辑链条的过程，使防护决策过程具有明显的表达形式。同时工作人员要创建起智能体决策依据的记录、保存和审计机制，对每一个交易风险研判的过程、决策依据以及处置结果都要加以详细地记载并加以保存，从而保证安全事件全过程可以被追溯并且能够提供证据支持，符合金融监管审查、责任判定和事件复盘等各方面的要求。

（三）搭建协同架构 强化全域联动

创建起统一调度、分层协作的多智能体防护体系，明晰不同区域、各类别、各个层次之间信息交流的标准和共享办法，冲破信息孤岛以及功能壁垒。设计出协同感知、协同研判、协同处置的工作流程来达到全网安全态势实时汇聚、统一分析、集中调度的目的[8]。对多智能体之间任务分发、指令同步以及资源调配等各方面加以改良，从而提高各个分布式节点间的协同程度和反应速度，并且加强整个防护体系对于大面积复杂袭击的综合抗风险能力。

以“网御星云安星人工智能安全运营项目”为例，在设计时要依照银狐木马这样的高阶威胁防护需求来创建一个统一体系中各个智能体之间的协作以及数据处理的模型。工作人员要形成规范的行为感知、AI 研判、剧本响应、威胁狩猎这四个智能体之间跨域信息交互的标准和数据共享的机制，消除终端防护、网络流量分析、威胁情报等各个系统之间的信息孤岛以及功能壁垒。工作人员需要制定出一套完整的多智能体协同感知、协同研判、协同处置工作流程，把行为感知智能体串联起来形成攻击链条，创建起完整的威胁全景图，AI 研判智能体消除告警噪声做到秒级精

确判断，剧本响应智能体完成自动闭环处理，威胁狩猎智能体寻找隐藏的威胁。

（四）健全内生安全 夯实防护底座

以智能体全生命周期为依托创建起安全防护体系，在模型训练、数据采集、指令传递、运行部署这些重要环节上采取风险控制手段来加强智能体自身的抵抗攻击、抵御干扰和防范篡改的能力。加强数据全生命周期的加密以及权限精细化控制来防止数据被泄露、受到污染、被人窃取的安全隐患，并且保证智能体所依靠的主要资源的安全和可靠。创建对抗样本、模型窃取、指令劫持等攻击的防御体系，使用异常检测、行为审计、故障隔离和动态恢复的办法来加强系统的抗风险能力[9]。

在“紫金山实验室无人系统内生安全智能体研发”项目中，工作人员要按照无人系统（如无人驾驶汽车、无人机）的安全防护要求来创建包含 AI 智能体整个生命周期的内生安全防护体系。工作人员需要对智能体模型训练过程中所用到的数据做全流程加密以及权限精细化管理，防止出现数据泄漏、污染等情况发生，保证智能体所依靠的主要数据是安全可靠的，在智能体指令传输时使用加密传输协议来防止被攻击者实施指令劫持或者篡改行为，从而保证指令传输过程中的完整性和安全性，在智能体运行部署期间创建起对抗样本、模型窃取等攻击的防护体系，借助异常监测、行为审计、故障隔离和动态恢复措施，加强智能体自身的抗攻击、抗干扰水平。同时工作人员要创建起智能体自身的安全评价体系，定时展开漏洞检测，渗透测试以及安全性检验工作，迅速修补可能存在的薄弱环节，从而保证 AI 智能体可以应对由 AI 辅助引发的快速变种威胁[10]。

三、结语

综上所述，本文主要研究了智能体驱动的网络安全和数据安全防护的研究背景、意义、主要难点以及具体的实现路径，通过四个核心策略结合真实的项目案例进行说明，在场景建模、可解释性框架建立、协同架构搭建、内生安全防护等各方面都给出了具体的实施方法。研究确定了智能体破解传统防护短板、提高防护智能化水平的核心作用，整理出目前防护工作中存在的主要问题和解决办法，用案例证明了所设计的防护路径是可行的、实用的，给智能体在网络安全、数据安全领域规范化的、规模化应用提供理论依据和方向指导，促进网络安全、数据安全防护体系的更新换代，适应数字化转型的安全防护与合规要求。

参考文献

[1] 张文雨，徐勇，孙健，等 . 网络攻击下多智能体系统攻击检测设计与分布式弹性控制 [J]. 自动化学报 ,2025,51(10):2347-2358.

[2] 苏德悦 . 人工智能发展进入智能体时代网络安全又迎新挑战 [J]. 人民邮电报 ,2025-08-11(09).

[3] 亚马逊云开发者 .AgenticAI 应用的隐私和安全 [J/OL]. 稀土掘金 ,2025-11-14.https://juejin.cn/post/7572028313930727464.

[4] 中国科学技术大学，NUS.G-Safeguard:ATopology-GuidedSecurityLensandTreatmentonLLM-basedMulti-agentSystems[C].ACL'25,2025.

[5] 李娟，王浩，张磊 .AI 智能体在网络安全防御中的场景适配与泛化能力优化研究 [J]. 网络安全技术与应用 ,2024,(8):45-48.

[6] 赵志国 . 人工智能与安全产业深度融合的路径与实践 [J]. 中国工信新闻网 ,2025-08-11(01).

[7] 陈明，李丽，张强 . 可解释 AI 智能体在金融数据安全防护中的应用研究 [J]. 金融科技时代 ,2024,(10):78-82.

[8] 周鸿祎 . 智能体时代网络攻防体系的重构与升级 [J]. 互联网安全大会论文集 ,2025,(1):1-6.

[9] 王丽，刘阳，陈浩 . 多智能体协同防护在企业网络安全中的应用实践 [J]. 计算机应用研究 ,2024,41(7):2108-2112.

[10] 张凯，李娜，王鹏 . 智能体内生安全防护体系构建与应用研究 [J]. 计算机工程与应用 ,2023,59(12):123-129.

