中国科学 : 信息科学 2025 年 第 55 卷 第 11 期: 2643–2644
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
人工智能应用中的数据安全专刊

编者按

人工智能技术的迅猛发展正深刻重塑人类社会的生产方式、认知体系与产业格局. 随着大模型、生成式人工智能和智能决策系统的持续演进, 人工智能的应用边界不断拓展, 其安全性与可信性问题也日益成为全球关注的焦点. 人工智能安全不仅涉及算法与模型的鲁棒性、隐私保护等基础问题, 还贯穿数据采集、模型训练、系统部署与决策治理的全生命周期安全保障. 在持续提升智能水平与应用效能的同时, 如何有效防范隐私泄露、对抗攻击、模型滥用与内容失真等风险, 已成为学术界与产业界亟待解决的重大挑战.

为系统呈现人工智能安全领域的最新研究进展与创新成果,《中国科学: 信息科学》特策划出版了“人工智能应用中的数据安全专刊”. 本专刊聚焦人工智能系统的安全理论与工程实践, 内容涵盖联邦学习安全、对抗性攻击方法、隐私保护与防御技术以及系统与评估框架等多个方向, 展示了我国科研团队在人工智能可信计算与安全防护方向的多维探索. 本专刊共遴选 16 篇高水平研究论文, 系统反映了人工智能安全从基础理论、算法机制到系统验证与评估体系的最新进展.

联邦学习以“数据不出域、模型协同训练”为核心理念, 在隐私保护型机器学习中发挥着重要作用. 然而, 在非独立同分布数据、拜占庭节点及不可信参与方等复杂环境下, 其安全性面临严峻挑战. 田晖等在“SDFL: 一种隐私保护和拜占庭鲁棒的去中心化联邦学习方案”中提出去中心化联邦学习安全框架, 通过跨节点梯度一致性验证与多层签名聚合机制实现无中心条件下的鲁棒聚合, 并结合局部模型可信度评估提升全局收敛稳定性. 张沁楠等在“自适应拜占庭鲁棒的差分隐私联邦学习”中融合差分隐私与拜占庭容错聚合机制, 设计自适应噪声注入与动态聚合策略, 可依据训练收敛速率实时调整开销, 在抗攻击能力与模型精度之间取得最佳平衡. 陈星星等在“一种适用于无辅助计算车载网络的联邦学习隐私保护方案”中针对车载协同计算场景提出轻量级加密聚合与局部扰动保护机制, 在资源受限网络中有效抵御中间人攻击与模型逆推风险. 禹勇等在“基于更新残差的差分隐私联邦遗忘学习机制”中构建基于残差更新的差分隐私遗忘算法, 利用梯度溯源机制实现特定样本贡献的可验证消除, 满足数据合规场景下的“安全可遗忘”要求. 范青等在“基于模型分解与加权聚合的联邦元遗忘”中结合联邦元学习与模型分解技术, 提出层级加权聚合算法, 通过任务相关性自适应调度与模型分解校准, 有效减弱历史样本残留效应, 提升动态任务切换下的可控性与安全性.

对抗性攻击揭示了深度学习系统在面对微小扰动时的脆弱性, 是人工智能安全研究的核心内容之一. 本专刊在攻击建模、侧信道风险与防御评测方面形成了多层次创新. 陈淑红等在“博弈论驱动的多层次扰动集成对抗攻击”中基于博弈论驱动的多层扰动策略, 提出可自适应调整扰动幅度与方向的多层次对抗攻击模型, 利用层间协同扰动机制提升攻击的迁移性与隐蔽性, 为鲁棒性研究提供理论工具. 李昊等在“DMS-MIA: 面向 TEE 保护机器学习模型的磁盘重放与多指标序列成员推理攻击”中关注可信执行环境 (TEE) 下潜在漏洞, 提出基于磁盘重放与多指标序列融合的成员推理攻击方法, 系统揭示 TEE 封装机制在访问模式侧信道下的隐私暴露风险. 宋亚飞等在“基于改进知识蒸馏的黑盒攻击方法”中从知识蒸馏与模型压缩视角设计改进的黑盒攻击框架, 通过代理模型蒸馏增强攻击可迁

引用格式: 黄欣沂, 何德彪. 人工智能应用中的数据安全专刊编者按. 中国科学: 信息科学, 2025, 55: 2643–2644, doi: 10.1360/SSI-2025-0442

移性, 使模型鲁棒性测试更贴近真实威胁场景. 张国明等在 “无需触发器与辅助数据集的模型后门攻击” 中提出无需触发器与辅助数据集的隐式后门注入机制, 利用生成扰动空间中的语义依附特征实现潜隐式后门嵌入, 揭示当前模型认证机制的潜在漏洞. 这些研究共同完善了人工智能安全威胁模型的体系化刻画, 为构建多层次防御策略提供理论依据.

隐私保护构筑了人工智能安全的底层防线. 本专刊展示了从密态计算、图神经网络防御到边缘智能隐私检测的多层次研究. 李洪伟等在 “支持双外包的轻量级函数隐藏内积加密方案” 中提出支持双外包的函数隐藏内积加密方案, 通过在云端外包解密与密钥生成计算, 实现对模型参数与输入数据的全程保护, 显著降低用户端的计算负担, 为数据评估与安全外包提供实用方案. 邓贤君等在 “Bandage: 针对图对抗攻击的双向嵌套的防御算法” 中针对图神经网络的对抗攻击问题, 构建双向嵌套防御算法, 通过在图结构与嵌入空间中同时引入双通道约束, 有效抵御节点扰动与结构攻击. 高海昌等在 “训练 – 推理协同的图神经网络成员推理防御框架” 中提出训练 – 推理协同防御框架, 在训练阶段动态嵌入隐私约束, 并在推理阶段结合置信度重标定与差分检测, 形成闭环防御体系, 显著降低成员推理成功率. 王滨等在 “基于并行深度神经网络的物联网边缘侧隐私数据检测” 中聚焦物联网边缘计算环境, 提出基于并行深度神经网络的隐私数据实时检测框架, 可在低延时条件下对数据流进行分层识别与自动隔离, 提升边缘节点的隐私感知与防护能力.

人工智能安全的研究不仅依赖算法创新, 还需要可度量、可验证的系统评估框架. 刘玉岭等在 “从编译到反编译: 基于源码级转换的高效水印去除方法” 中提出基于源码级转换的模型与代码水印去除机制, 在编译与反编译阶段引入语义一致性映射, 实现对 AI 模型水印的高效分析与去除, 为知识产权保护与可信模型鉴定提供新途径. 丁旭阳等在 “基于神经网络决策边界与一致性分析的对抗样本攻击检测方法” 中基于神经网络决策边界一致性分析构建对抗样本检测体系, 通过多层特征空间映射与判别一致性指标实现对抗扰动的精准识别, 显著提升检测的普适性与泛化能力. 杨伟平等在 “M3-SafetyBench: 多领域多场景多维度的大语言模型安全评估体系” 中提出的 M3-SafetyBench 多维度评测框架覆盖内容安全、鲁棒性与可控性等多个维度, 建立统一的测试基准与评测指标体系, 支持对多模态与大语言模型进行跨领域安全评估, 为 AI 系统监管与治理提供可量化工具.

最后, 诚挚感谢参与本专刊的所有作者、审稿专家和编辑同仁的辛勤付出. 正是他们的专业精神与协作努力, 使本专刊能够高质量地呈现人工智能安全领域的研究成果, 为相关领域的持续研究与学术交流提供了有益支持.

特约编辑:　黄欣沂　暨南大学
何德彪　武汉大学