中国科学 : 信息科学 2026 年 第 56 卷 第 2 期: 481–482
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
精选文章导读

# 人工智能安全与隐私综述

何新磊[1], 徐国文[2], 韩星烁[3], 王骞[4], 赵令辰[4], 沈超[5], 蔺琛皓[5], 赵正宇[5], 李前[5], 杨乐[6], 纪守领[7], 李少锋[8], 朱浩瑾[9], 王志波[10], 郑锐[10], 朱天清[11], 李琦[12], 贺超翔[13], 王啟帆[14], 胡宏盛[15], 王烁[13], 孙士锋[13], 姚宏伟[10], 秦湛[10], 陈恺[16], 赵月[16], 李洪伟[2], 黄欣沂[17*], 冯登国[18*]

1. 香港科技大学 (广州) 信息枢纽, 广州 511453, 中国
2. 电子科技大学计算机科学与工程学院, 成都 611731, 中国
3. College of Computing and Data Science, Nanyang Technological University, Singapore 639798, Singapore
4. 武汉大学网络空间安全学院, 武汉 430072, 中国
5. 西安交通大学网络空间安全学院, 西安 710049, 中国
6. 西安交通大学信息与通信工程学院, 西安 710049, 中国
7. 浙江大学计算机科学与技术学院, 杭州 310058, 中国
8. 东南大学计算机科学与工程学院, 南京 211189, 中国
9. 上海交通大学计算机科学与工程系, 上海 200240, 中国
10. 浙江大学区块链与数据安全国家重点实验室, 杭州 310058, 中国
11. 澳门城市大学数据科学学院, 澳门 999078, 中国
12. 清华大学网络科学与网络空间研究院, 北京 100086, 中国
13. 上海交通大学网络空间安全学院, 上海 200240, 中国
14. School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK
15. School of Information and Physical Sciences, University of Newcastle, Newcastle 2287, Australia
16. 中国科学院信息工程研究所, 北京 100093, 中国
17. 暨南大学网络空间安全学院, 广州 511443, 中国
18. 密码学国家重点实验室, 北京 100878, 中国
* 通信作者. E-mail: xyhuang81@jnu.edu.cn, fengdg@263.net

随着人工智能 (artificial intelligence, AI) 技术从传统机器学习不断迈向生成式智能与基座模型 (foundation models), 其在金融、医疗、交通、工业控制等关键领域的渗透日益加深, 系统形态也愈发复杂. AI 系统已经从单一模型、单节点训练的集中式形态, 演化为涵盖联邦学习 (federated learning)、拆分学习 (split learning)、云边协同以及多模态基座模型在内的庞大生态. 在这一过程中, 安全与隐私问题正从边缘议题演变为制约 AI 可靠落地的核心瓶颈: 数据投毒、后门植入、模型劫持、梯度反演、对抗样本与模型窃取等攻击手段不断涌现 [1,2], 大型语言模型 (large language models, LLMs) 与生成式模型还面临越狱 (jailbreak)、提示注入 (prompt injection)、训练数据泄露等新型风险. 如何在充分释放人工智能潜能的同时, 有效防范贯穿全生命周期的安全威胁与隐私泄露, 已成为亟需系统化梳理和深入研究的关键课题.

*SCIENCE CHINA Information Sciences* 在 2025 年 68 卷第 8 期出版了冯登国院士等的综述文章 “Artificial intelligence security and privacy: a survey”, 本文系统回顾了近年人工智能安全与隐私领域的大量代表性研究工作, 从 “训练 – 推理” 两个阶段、“集中式 – 分布式” 两类部署形态以及 “传统模型 – 基座模型” 两类技术

**英文原文:** He X L, Xu G W, Han X S, et al. Artificial intelligence security and privacy: a survey. Sci China Inf Sci, 2025, 68: 181101, doi: 10.1007/s11432-025-4388-5

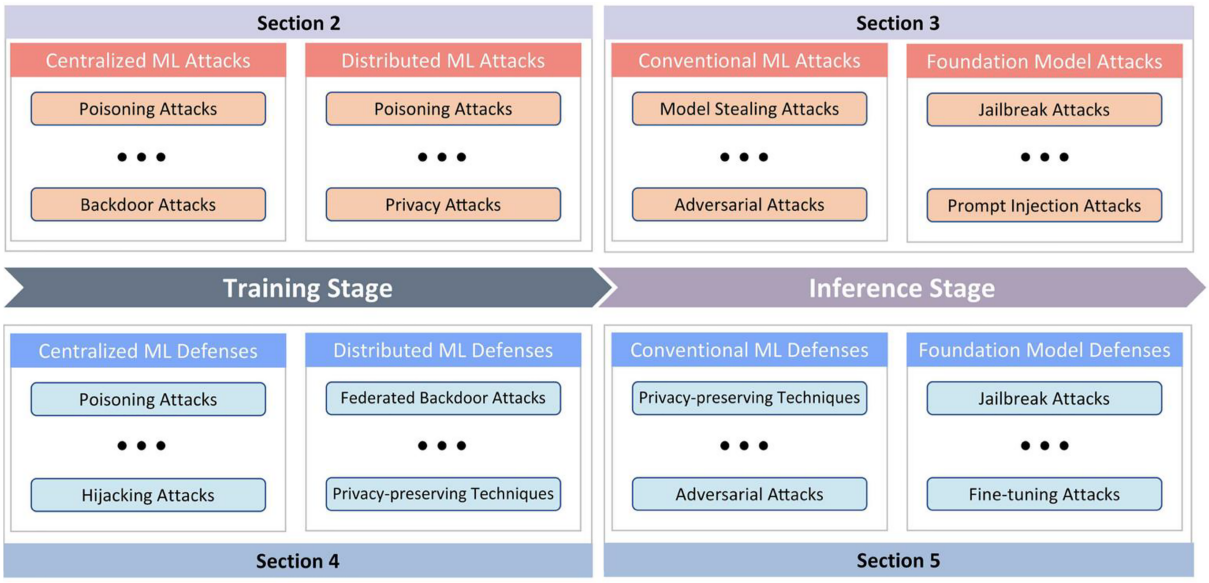


**图 1**　(网络版彩图) 人工智能模型中的攻击与防御分类. **ML: machine learning.**

**Figure 1**　(Color online) Taxonomy of attacks and defenses in AI models. ML stands for machine learning.

范式三个维度, 对 AI 系统中的安全威胁与防护机制进行了细致的分类与分析 (如图 1 所示). 在训练阶段, 重点总结了非定向数据投毒、后门攻击与模型劫持等针对集中式训练流程的攻击方式, 以及联邦学习与拆分学习场景下的模型投毒、数据投毒、梯度/特征反演和服务端操纵等新型威胁. 在推理阶段, 系统梳理了成员推断 (membership inference)、模型反演 (model inversion) 等隐私推断攻击, 故障注入与侧信道攻击, 对抗样本攻击及其物理实现, 以及通过大规模查询实现的模型窃取 (model stealing) 等问题, 并进一步扩展到基座模型和大模型面临的越狱、提示注入、恶意微调与安全对齐退化等前沿攻击形态, 呈现出一个从底层数据、模型参数到上层应用与生态的立体化攻防图景.

在防御层面, 以体系化视角总结了多层次的安全加固思路: 在训练阶段, 涵盖鲁棒优化与稳健训练、异常样本与恶意更新检测、后门分析与模型净化、差分隐私 (differential privacy) 与安全聚合 (secure aggregation) 等隐私保护训练机制, 以及面向联邦/拆分学习协议的安全设计与可信执行环境; 在推理阶段, 则包括对抗训练与认证鲁棒性方法、访问控制与水印技术、同态加密与多方安全计算等隐私增强推理方案, 以及针对大模型的安全对齐训练、安全评测与红队测试、提示过滤与日志审计等治理框架. 通过对典型攻击与防御方法的对比分析, 强调当前尚不存在 “一劳永逸” 的通用防御方案, 实际部署中必须在安全性、性能开销、系统复杂度与可用性之间进行精细权衡, 并根据应用场景构建分层、多点联动的纵深防御体系.

在此基础上, 本文进一步凝练出若干具有代表性的未来研究方向: 一是面向持续涌现的新型 AI 范式 (如多模态基座模型、代理化系统、联邦大模型等) 的安全与隐私威胁识别与专门防御设计; 二是从受控实验环境走向大规模、异构、动态真实系统的 “更现实” 的防御策略与评测基准; 三是将可解释性 (explainability) 与可理解性 (interpretability) 深度融入安全防护过程, 使模型能够解释对抗输入与异常行为, 从而提升关键场景中的可信度; 四是面向云、边、端协同部署环境的跨平台安全机制, 兼顾资源受限设备上的效率与鲁棒性; 五是进一步强化密码学、系统安全、软件工程、法律与伦理等多学科协同, 构建覆盖技术、制度与治理层面的综合性 AI 安全保障体系. 总体而言, 本文旨在通过对人工智能安全与隐私攻防格局的系统梳理, 为研究人员提供了清晰的知识地图与问题脉络, 也为产业界设计和部署面向未来的安全可信 AI 系统提供参考与启示.

## 参考文献

1　Hu Y, Kuang W, Qin Z, et al. Artificial intelligence security: threats and countermeasures. ACM Comput Surv, 2021, 55: 1–36

2　Rahman M M, Arshi A S, Hasan M M, et al. Security risk and attacks in AI: a survey of security and privacy. In: Proceedings of IEEE Annual Computers, Software, and Applications Conference, 2023. 1834–1839