人工智能安全技术与发展研究

李 可 高嘉阳 武 凯
北京计算机技术及应用研究所 北京 海淀区 100854

摘 要：人工智能安全技术是保障AI系统全生命周期稳定运行的核心支撑，涵盖算法、数据、系统三大维度安全防护。当前技术发展聚焦对抗样本防御、数据隐私保护及模型鲁棒性提升，同时面临安全与性能平衡、跨领域威胁识别等挑战。未来趋势呈现技术融合化、工具自动化及场景轻量化特征，需构建全球协同治理框架，推动AI安全技术向主动防御、全周期管控方向演进。

关键词：人工智能；安全技术；发展

引言：人工智能正加速渗透至社会各领域，但其安全风险如算法偏见、数据泄露、对抗攻击等日益成为技术应用的掣肘。从自动驾驶的决策安全到医疗AI的诊断可靠性，安全防护已贯穿AI全生命周期。本文聚焦人工智能安全技术的基础理论、前沿突破与发展挑战，探讨技术融合、政策协同与产业实践的协同路径，旨在为构建安全可信的AI生态提供理论支撑与实践参考。

1 人工智能安全技术基础理论

1.1 人工智能安全的核心内涵

（1）定义与范畴：人工智能安全是保障AI系统全生命周期稳定可靠运行、防范各类安全风险的技术与理论体系，核心范畴涵盖算法安全、数据安全、系统安全三大维度。算法安全聚焦模型训练与推理过程中的逻辑漏洞、偏见风险等；数据安全围绕训练数据的采集、存储、使用全流程，防范数据泄露、篡改等问题；系统安全则关注AI部署载体的硬件、软件及网络环境安全。（2）与传统信息安全的区别与联系：联系在于均以保障信息系统安全稳定为核心目标，共享加密、访问控制等基础安全技术。区别在于传统信息安全侧重防御外部攻击，而AI安全需应对模型自身缺陷、数据污染等内生风险；且AI的自主性、学习性使风险具有动态演化特性，防护难度更高[1]。

1.2 AI安全威胁模型

（1）攻击面分析：主要包括数据层、模型层与应用层。数据层面临数据投毒攻击，即攻击者篡改训练数据影响模型输出；模型层存在模型窃取、模型规避等威胁，攻击者通过查询获取模型参数或诱导模型误判；应用层易受对抗样本攻击，通过细微调整输入误导模型决策。（2）典型攻击场景与防御需求：典型场景包括智能驾驶中的对抗样本干扰、推荐系统的数据投毒舞弊等。防御需满足实时性要求，能动态识别未知威胁；同时需兼顾AI系统性能，避免过度防护导致精度下降。

1.3 安全技术分类体系

（1）主动防御技术：通过鲁棒性训练提升模型对抗干扰能力，利用加密计算（如同态加密）保护数据与模型隐私，从源头降低风险。（2）被动检测技术：借助异常监测系统识别数据或模型的异常行为，通过模型审计定期排查算法漏洞与偏见，实现风险早发现。（3）隐私保护技术：基于联邦学习实现多方数据“数据不动模型动”，通过差分隐私在数据处理中加入噪声，保护用户隐私不泄露。

2 人工智能安全技术分析

2.1 对抗样本防御技术

（1）对抗样本生成原理与攻击类型：对抗样本通过在合法输入中添加人类难以察觉的细微扰动，利用模型决策边界的脆弱性诱导其产生错误判断，核心原理是基于梯度下降算法寻找最优扰动方向。攻击类型主要分为白盒攻击（攻击者知晓模型结构与参数）、黑盒攻击（仅知晓模型输入输出）及自适应攻击（根据防御机制动态调整攻击策略），其中黑盒攻击因适用性广，在实际场景中威胁性更强。（2）防御策略：输入预处理通过去噪、归一化等手段过滤扰动，降低对抗样本的攻击性；模型鲁棒性优化采用对抗训练、正则化等方法，拓展模型决策边界，提升对扰动的容忍度；检测机制通过构建异常检测模型，识别输入数据的分布偏差，精准区分对抗样本与合法输入。

2.2 数据隐私保护技术

（1）数据匿名化与加密传输：数据匿名化通过泛化、屏蔽、置换等技术去除身份证号、手机号等敏感信息，降低数据关联识别风险，同时需平衡匿名化程度与数据可用性；加密传输依托SSL/TLS、IPSec等协议构建安全传输链路，对数据在采集、传输环节进行端到端加密，防

范数据被窃听、篡改，保障数据流转安全。（2）联邦学习与分布式AI架构：联邦学习构建“数据不动模型动”的分布式训练模式，通过加密聚合各参与方模型参数，实现数据隐私保护与协同建模的双赢，按数据分布可分为横向（样本不同特征相同）、纵向（特征不同样本相同）及迁移联邦学习三类；分布式AI架构通过多节点协同计算，避免原始数据集中存储，减少单点泄露风险，同时提升模型训练的效率与容错性[2]。（3）差分隐私在AI训练中的应用：核心是在训练数据或模型梯度中添加可控的Laplace或Gaussian噪声，使攻击者无法通过模型输出反推单个样本信息，其核心指标“隐私预算”决定了隐私保护强度与模型精度的平衡。在实际应用中，通过自适应噪声调节、隐私预算分配优化等技术，在满足严格隐私要求的前提下，最大限度维持模型的泛化能力。

2.3 模型安全与可解释性

（1）模型窃取与逆向工程防御：针对模型窃取，采用输入输出扰动、查询频率限制、模型水印嵌入等技术，增加攻击者获取有效参数的难度与成本；防御逆向工程需从部署环节优化，通过模型量化、剪枝、混淆等手段隐藏核心结构，同时建立严格的访问控制与身份认证机制，限制非授权用户对模型的调用与查询。（2）可解释性AI（XAI）在安全审计中的作用：XAI通过特征归因、可视化热力图、逻辑规则提取等手段，清晰呈现模型决策的依据与过程，解决传统AI“黑箱”问题。在安全审计中，可解释性结果能帮助审计人员精准定位模型偏见、逻辑漏洞及潜在安全风险，为模型合规性验证、风险评估提供量化依据，提升审计效率与可信度。

2.4 AI系统供应链安全

（1）开源框架漏洞与依赖管理：TensorFlow、PyTorch等开源框架因代码开放性易引入漏洞，需建立常态化漏洞扫描、预警与更新机制；通过依赖项审计工具（如OWASPDependency-Check）全面排查第三方组件的安全风险，建立组件白名单制度，优先选用经过安全验证的版本，同时加强开源社区协作，推动漏洞快速修复。（2）硬件级安全保障（TPU/AI芯片安全）：从芯片设计、生产到部署全流程融入安全机制，采用硬件加密、可信执行环境（TEE）、物理不可克隆函数（PUF）等技术，保护芯片内存储的模型参数与敏感数据；加强芯片供应链管控，防范硬件被篡改、植入恶意组件，通过硬件级安全认证，筑牢AI系统的底层安全底座[3]。

2.5 法律与伦理安全框架

（1）AI伦理准则与合规性要求：基于公平、透明、问责、普惠等核心原则制定伦理准则，规范AI在医疗、司法、教育等关键领域的应用，防范算法偏见、歧视等问题；合规性要求需严格贴合《生成式人工智能服务管理暂行办法》《数据安全法》等法律法规，明确数据采集、模型训练、应用部署各环节的合规边界。（2）责任归属与监管机制设计：建立“研发-部署-使用”全链条责任归属体系，明确企业、研发人员、使用者在AI安全事件中的责任划分；构建分级分类监管机制，对自动驾驶、医疗诊断等高风险AI应用实施重点监管，通过动态监测、定期审计强化风险防控；同时建立投诉举报与纠纷解决机制，保障公众权益，推动AI技术安全合规发展。

3 人工智能安全技术发展趋势

3.1 技术融合趋势

（1）AI与区块链结合的分布式安全架构：二者融合将构建去中心化的可信AI生态，依托区块链的共识机制与不可篡改特性，解决分布式AI训练中的数据可信度与节点信任问题。通过零知识证明实现计算正确性验证，借助智能合约规范节点行为，在保障数据隐私的同时降低单点故障风险，为分布式GPU网络等场景提供安全支撑。（2）量子计算对AI安全的挑战与机遇：挑战在于量子计算可破解现有加密体系，威胁AI模型参数与敏感数据安全；机遇则体现在量子计算能提升AI模型训练效率，同时催生出后量子密码等新型防护技术，推动AI安全防护体系向更高强度演进。

3.2 自动化安全工具链

（1）基于AI的自动化漏洞挖掘与修复：借助大语言模型与图神经网络，实现漏洞特征的精准提取与复杂业务逻辑的深度覆盖，通过静动协同框架提升漏洞检测精度并降低误报率。同时，利用生成式AI自动生成修复代码，形成“检测-评估-修复-验证”的全流程自动化闭环，大幅提升漏洞治理效率。（2）安全运维（SecOps）的智能化升级：将AI技术融入运维全流程，通过实时监测系统日志与行为数据，实现安全威胁的提前预警与快速响应。依托智能决策引擎自动调度防护资源，推动SecOps从被动响应向主动防御转型，提升复杂AI系统的运维安全性与效率[4]。

3.3 新型应用场景安全需求

（1）自动驾驶、医疗AI等高风险领域的安全标准：此类领域正推动建立全生命周期安全认证标准，量化模型鲁棒性、决策可靠性等指标。例如自动驾驶需通过数百种复杂场景测试保障决策准确率，医疗AI需规范数据脱敏与诊断可解释性要求，以降低技术应用风险。（2）边缘计算与物联网环境下的轻量化安全方案：针对边缘设备算力有限的特点，发展轻量化加密算法、精简版鲁

棒性模型等技术。通过本地数据处理减少传输风险，构建“端-边-云”协同的分级防护体系，满足物联网场景下AI应用的安全需求。

3.4 全球治理与标准化进程

（1）国际组织（IEEE、ISO）的AI安全标准制定：IEEE、ISO等正加速推进跨领域AI安全标准构建，覆盖模型安全评估、数据隐私保护等核心维度，旨在解决现有规范碎片化、操作性弱的问题，推动形成统一的技术安全认证体系。（2）跨国技术合作与政策协同机制：各国正加强AI安全领域的跨国协作，通过共享漏洞信息、联合开展技术验证等方式应对全球性安全挑战。同时推动政策协同，对齐高风险AI应用的合规要求，构建兼顾技术创新与风险防控的全球治理框架。

4 人工智能安全的挑战与对策建议

4.1 主要技术挑战

（1）安全与性能的平衡难题：AI安全防护措施往往会增加系统算力开销、降低模型推理效率。例如鲁棒性训练会提升模型对抗能力，但可能导致精度下降；加密计算能保障数据隐私，却会延长处理耗时，如何在强化安全防护的同时维持AI系统的核心性能，成为技术落地的关键瓶颈。（2）跨领域安全威胁的复杂性：AI技术广泛渗透于自动驾驶、医疗、金融等多领域，不同领域的应用场景、数据类型差异显著，导致安全威胁呈现跨域传导、复合叠加的特征。单一领域的防护方案难以应对跨领域风险，增加了威胁识别与防御的难度。（3）长期安全演化的不可预测性：AI模型具有自主学习与迭代能力，随着应用场景拓展和数据迭代更新，可能衍生出新的安全漏洞。同时，攻击者的技术手段也在持续升级，使得AI安全风险呈现动态演化特征，长期防护的预判与应对难度极大[5]。

4.2 非技术挑战

（1）法律滞后性与监管空白：AI技术发展速度远超法律修订进程，针对生成式AI、自动驾驶等新型应用的安全责任界定、侵权认定等法律条款尚不健全，监管机制存在覆盖盲区，难以有效规范技术应用中的安全行为。（2）公众认知偏差与信任危机：公众对AI安全风险的认知存在两极化，部分群体过度担忧技术风险，而部分群体则忽视潜在隐患。同时，AI安全事件的频发易引发公众对技术的信任危机，阻碍AI技术的良性推广。（3）人才缺口与学科交叉需求：AI安全领域需要兼具计算机科学、cybersecurity、法学等多学科知识的复合型人才，当前相关人才培养体系不完善，人才供给缺口较大，难以支撑技术研发与安全防护工作的全面推进。

4.3 对策建议

（1）技术层面：构建动态防御体系，结合AI技术实现威胁的实时监测与自适应响应；推动安全前置设计，将安全需求融入AI系统的研发、部署全生命周期，从源头降低安全风险，平衡安全防护与系统性能。（2）政策层面：加快完善AI安全相关法律法规，细化不同领域的安全标准与责任界定；建立全球协同治理机制，加强跨国政策协同与风险信息共享，填补监管空白，规范技术应用边界。（3）产业层面：加强产学研协同创新，推动高校、科研机构与企业联合开展技术研发与人才培养；培育AI安全服务生态，鼓励第三方机构提供安全检测、评估等专业服务，提升全产业的安全防护能力。

结束语

人工智能安全技术是驱动AI可持续发展的基石，其发展需兼顾技术创新与风险防控的双重使命。面向未来，技术融合、自动化工具链与轻量化方案将成为突破安全瓶颈的关键路径，而全球治理协作与跨学科人才培养则是构建长效防护体系的根本保障。唯有以动态防御思维应对不断演化的安全挑战，推动技术、政策与产业的协同共进，方能实现AI“向善而行”的终极目标。

参考文献

[1]李建彬,谯婷,秦淑梅,等.人工智能安全综述[J].中国信息安全,2023(5):82-83.

[2]景慧昀,魏薇,周川,等.人工智能安全框架[J].计算机科学,2021(7):102-103.

[3]彭长根.人工智能安全治理挑战与对策[J].信息安全研究,2022(4):60-62.

[4]朱倩倩,唐志敏,王新哲.人工智能安全框架研究[J].工业信息安全,2022(10):138-139.

[5]尹刚.安全生产人工智能技术应用与发展[J].工程地质学,2025(6):39-41.