 DOI: https://doi.org/10.12349/iser.v6i11.8106

Discussion on security risks and protection measures of artificial intelligence large model

Junyong Jiang

Shanghai Digital Security Technology Co., Ltd., Shanghai, 200000, China

Abstract

Artificial intelligence (AI) large models, serving as the core driving force of next-generation intelligent systems, have demonstrated significant value across diverse fields including natural language processing, computer vision, and automated decision-making. However, during their widespread application, these AI models continue to face complex and multifaceted security risks. While threatening the stability and controllability of threat models, they also trigger governance dilemmas and social trust crises. To address these challenges, this paper first systematically analyzes the security risks confronting AI large models, then proposes scientifically feasible protective measures based on research and practical insights. The aim is to provide actionable guidance for advancing the optimal application and development of China's AI large models.

Keywords

AI large models; security risks; protective measures; exploration

人工智能大模型的安全风险及防护措施探讨

江均勇

上海数字安全科技有限公司，中国 · 上海 200000

摘　要

自然语言处理、计算机视觉以及自动决策等诸多领域，人工智能大模型身为新一代智能系统的核心驱动力在其中发挥出重要价值。然而在广泛应用进程里，人工智能大模型仍面临着复杂且多元化的安全风险，它们在威胁模型的稳定性与可控性的同时，还会引发治理困境与社会信任危机。本文为应对上述挑战，先就人工智能大模型所面临安全风险展开系统剖析，随后结合研究及实践情况下提出相对应的科学可行防护措施，希望能够为推动我国人工智能大模型更好地应用及发展提供一定帮助。

关键词

人工智能大模型；安全风险；防护措施；探讨

1 引言

随着深度学习技术的突破、算力的显著提升、海量数据的积累，人工智能大模型的发展日新月异。人工智能大模型具有更多的参数和更复杂的结构，可以更好地捕捉数据中的复杂关系和模式，从而提高模型的性能。人工智能大模型的快速发展为各行业带来了前所未有的机遇，但同时也带来了多样化的安全风险。这些风险贯穿于模型、数据和应用等各个层面，并伴随着相应的攻击手段。

2 人工智能大模型的安全风险分析

2.1 数据安全风险

大模型训练和推理过程中数据安全风险乃是首要且最为基础的安全隐患，表现为数据采集、存储、传输以及使用阶段面临的泄漏、篡改乃至窃取可能。人工智能大模型的超大规模训练语料及样本集，其中包含个人隐私信息、敏感行业数据甚至是国家关键信息基础设施相关资料，若数据缺乏严格脱敏与合法授权极有可能出现暴露和数据滥用情况[1]。数据安全风险危害体现在三个层面：个人层面容易出现隐私信息不可逆的泄露情况或是画像被滥用；企业层面会致使知识产权以及商业机密流失，造成市场竞争力降低；国家层面一旦涉及公共安全以及关键行业的数据泄露会使国家网络安全屏障遭到削弱且极有可能成为外部势力展开攻击的关键突破口。

2.2 模型攻击风险

所谓模型攻击风险主要指不法人员会利用人工智能大模型训练、推理进程里的漏洞采取包括对抗性攻击、模型窃取、反向推理或者后门注入等手段致使其性能与可靠性遭到破坏。受大模型复杂的参数空间与深度网络结构二者天然存

【作者简介】江均勇（1981–），男，中国重庆市人，硕士，从事云计算、大数据、网络安全、数据安全、人工智能及人工智能安全研究。

在高维脆弱性是模型攻击风险产生的根源，这导致其对于精心设计的扰动样本呈现出显著的易受攻击性。该风险危害体现在人工智能大模型预测精度降低及功能失效，同时不法人员还可能利用其制造虚假信息、操纵决策流程与规避安全检测，这会对如金融、医疗甚至政务等关键行业应用的稳定运行构成威胁，并引发社会信任的系统性风险。

2.3 算法偏见与不确定性

人工智能大模型在学习与推理时因训练样本分布不均衡、存在数据噪声、算法机制有缺陷等原因致使输出结果出现系统性偏差或不确定性过高，这即为算法偏见与不确定性风险。该风险的危害主要体现在两个方面：算法偏见有可能致使公共服务、金融信贷以及招聘等领域出现群体性的不公情况，这会让社会不平等以及公众不信任的状况加剧 [2]。模型输出结果不确定性会导致其可靠性大幅降低，这造成如自动驾驶、医疗辅助诊断这类安全关键场景里出现误判或漏判，继而对生命财产安全构成严重威胁。

2.4 伦理与合规风险

人工智能大模型的研发、部署以及应用进程里，因欠缺伦理约束且未能与法律规范良好匹配进而引发的风险被称作伦理与合规风险，其产生原因在于强大的生成与推理能力。比如，人工智能大模型能在未经授权时生成虚假文本，伪造图像或合成语音，随后被用于虚假舆论操纵和信息欺诈。其危害表现在三个层面：用户层面会侵犯其合法权益与个人自由；市场层面容易诱发不正当竞争以及技术垄断；社会治理层面一旦人工智能大模型被肆意用于虚假信息的扩散以及社会情绪的操纵，那么社会秩序与国家安全将会遭受严重冲击。

3 人工智能大模型的防护措施

3.1 数据治理与隐私保护

人工智能大模型安全治理体系里数据治理和隐私保护占据着核心地位，其施行措施涉及数据采集、模型训练与存储传输等环节。数据采集阶段需构建基于分级授权的审查体系，严格分类管理大模型采集源敏感级别以及语料类型，同时运用伪造化语料替换、k- 匿名化处理、可验证的数据脱敏算法等技术保证任何单一标识符都无法达成用户身份的重识别。对于跨境数据流动，应依托国家网络安全审查要求构建跨境数据交换协议，此协议基于可信执行环境（TEE）同态加密与多方安全计算（MPC），随后再借助密钥分片与门限签名实现跨境使用的合规可控。大模型训练的环节当中差分隐私机制应当被全面引入，借助添加噪声扰动的方式达到在梯度上传过程里对敏感信息的保护，并结合联邦学习架构把训练数据留存于本地节点，只针对模型参数层面开展加密聚合，同时借助同态加密技术保障训练过程的不可见性。此外，对大模型训练过程实施实时监测与隔离，这可通过部署基于可信硬件的隔离执行环境予以实现，以避免训练数据不会因内存溢出或恶意调试而出现外泄情况 [3]。数据存储传输环节中采用分布式加密存储架构，把核心数据按分片化形式，各自加密存储到不同物理节点上，同时将抗量子密码算法与端到端加密传输协议相结合，以此降低数据于传输链路里的拦截以及解密风险。最后，借助基于联盟链的溯源系统记录数据访问与修改日志，并采取智能合约达到对访问请求的自动化审计与动态校验，以此规避篡改行为与越权操作。

3.2 模型安全加固

模型安全加固是人工智能大模型应对复杂威胁的关键组成，其要点在于借助训练、推理、参数保护与部署等措施全方位提高模型的鲁棒性与抗攻击能力。大模型训练阶段运用对抗性训练模式，简单而言样本里引入多样形式的对抗扰动，让模型能够在黑盒查询攻击、梯度可迁移性攻击之下维持稳定收敛。随后基于大规模数据和特征空间扰动两方面增强手段，提升大模型在非独立同分布场景当中的泛化性能。同时优化进程里应引入随机梯度扰动、Dropou 等随机化正则化手段，以此来破坏攻击者借助梯度信息构建精确扰动的可能性。大模型推理阶段输入检测机制需要建立起来，借助统计特征分析、傅里叶谱域分析以及基于生成对抗网络的输入分布一致性检测对输入样本开展多层次的安全筛查工作，从而对抗性样本、规避式攻击以及指令注入类攻击都能有效规避。大模型参数保护阶段可针对云端推理与分布式训练两个环节引入同态加密与安全多方计算，以实现模型参数在整个计算进程中始终处于加密状态，从而防止由于 API 调用频次推断或者梯度泄露分析而引发的模型结构与权重被窃取的情况。随后利用差分隐私机制在梯度更新过程当中添加噪声用来削弱攻击者凭借反演技术恢复训练数据的能力，并且在针对后门攻击防护则构建模型完整性验证与参数审计机制。另外，借助权重稀疏化可视化办法、隐藏触发器激活模式分析和梯度分布异常检测以对潜在后门结构展开剖析并予以剔除。大模型推理服务部署阶段当中应当采用一种将多模型冗余架构跟可信执行环境（TEE）相融合的防护体系，简而言之凭借不同架构模型之间的交叉验证以及硬件级别实现隔离，防止由于单点模型的失陷而致使系统性风险出现扩散的情况，同时还可引入远程证明机制保障部署环境具备不可篡改性以及可信性 [4]。

3.3 算法可解释性与公平性提升

由技术路径以及制度设计这两个层面构建起系统化的防护机制，以此提升人工智能大模型的算法可解释性与公平性。针对不同模型架构的可解释性环节，需引入多层次解释框架，具体为：一是局部解释层面借助 LIME 与 SHAP 等办法针对单个预测结果展开特征贡献分解工作，联合梯度反传、特征遮蔽等技术搭建细粒度的特征敏感度分析以此来揭露模型面对输入扰动时的响应规律；二是全局解释层面就基于 Transformer 大模型，采用特征空间聚类方法、信息

流路径追踪以及注意力权重可视化对模型整体的决策通路与参数分布进行建模，这样一来可提供数值化与可视化并行的解释框架。构建跨样本的解释性验证机制时要运用因果图建模、反事实生成等办法，以此防止解释仅仅停留在统计相关性却忽略了潜在的因果关系。公平性提升环节上集中于数据、算法以及结果三个层面推行分层治理举措：一是数据层面借助对抗式生成补全样本、多分布重采样，以及对群体敏感属性进行分层加权等方式减轻训练数据里群体特征不均衡所产生的影响；二是算法层面把加性正则化项、群体公平性约束和个体公平性约束一起嵌入到目标函数当中，在梯度更新阶段达成精度最优化和公平性约束的协同优化，同时再借助多目标优化框架来达成权衡参数的动态调节；三是结果层面要运用输出分布再映射、后验概率分段调整，以及分群体阈值校准的策略减小不同群体在模型输出错误率与置信度方面的系统性差异。另外，需要建立针对不确定性风险的基于贝叶斯神经网络与深度集合学习的区间估计机制，该机制要为每个预测结果生成概率分布以及置信区间，结合蒙特卡罗 Dropout 和 Bootstrap 重采样以提升不确定性度量的稳健性。在具体应用场景里针对如公共服务、司法裁决以及医疗诊断这类高敏感性行业要专门构建强制性可解释性披露机制与实施独立第三方审计制度，随后将差分隐私保护与联邦学习框架引入，以促使数据安全合规的情况下让算法可解释性与公平性得以提升，最终从多维度给人工智能大模型安全防护提供技术支撑。

3.4 制度监管与多方协同

制度监管和多方协同应当构建成多层次跨域的综合治理机制，以此来应对大模型潜在的合规与安全风险。国家法律层面除须针对大模型训练数据的来源、脱敏处理，跨境流动以及安全审查等明确具体标准保证数据链条全程可追溯外，还必须制定涉及数据安全、知识产权模型可解释性以及生成内容责任等方面的专项法规，同时实施动态修订机制来适应大模型技术的迭代。大模型开发与部署行业标准构建层面则须形成分层次的技术规范，具体涵盖算法透明度指标、可复现性验证流程、模型安全性评估以及内部审计制度等内容，同时以统一测试平台对模型性能、鲁棒性以及安全防护能力开展标准化验证 [5]。企业内部治理层面需强制组建人工智能伦理与合规审查委员会，并清晰界定其于模型生命周期里针对数据采集，模型训练推理应用以及迭代更新这一整个流程的监督职责，同时建立涵盖算法偏差检测、异常行为监控以及访问权限管理在内的多维度风险识别机制，借助日志追踪和责任归属机制达到违规事件的溯源与处置。社会协同层面则推动公众、第三方评估组织以及学术机构参与监督机制，借助构建可控试点机制与沙箱环境就新型大模型技术展开阶段性评估与受控部署，并依据多源反馈对模型行为实施迭代优化。另外，实施跨部门信息共享与预警机制，并在集中监管平台中纳入安全事件、风险评估及合规审查数据，以推动政府、产业与科研单位的实时联动。通过动态调整监管策略和技术防护措施，在创新迭代与风险控制之间形成系统化的协同运作模式，有效支撑大模型在复杂应用环境中的安全可控性。

4 结语

在科技创新与产业升级不断深入背景下，人工智能大模型的安全风险逐渐浮现且呈现出多维度的复杂性，为此上文经由对数据安全、模型攻击、算法偏见与不确定性、伦理与合规 4 种大模型面临的安全风险研究分别探讨数据治理与隐私保护、模型安全加固、算法可解释性与公平性提升，以及制度监管与多方协同这四方面的防护措施，从而给人工智能大模型的安全可控发展给予系统性路径。

参考文献

[1] 吕延辉,张博,高彦恺.基于人工智能安全治理框架的大模型系统安全防护研究[J].中国信息安全, 2024(10):38-41.

[2] 蔡佳,黄璇,童国炜,等.人工智能大模型安全现状研究[J].海峡科技与产业, 2025, 38(1):53-57.

[3] 史锋,张永晋,瞿崇晓,等.ChatGPT类大模型技术的安全风险与应对措施研究[C]//第十二届中国指挥控制大会.中电海康集团有限公司;中国电子科技集团公司第五十二研究所, 2024.

[4] 李立.基于人工智能的网络安全风险评估模型研究[J].黑龙江科技信息, 2021, 000(007):103-104.

[5] 邱惠君,张瑶.大模型发展对人工智能安全风险治理的挑战和应对分析[J].工业信息安全, 2023(2):65-72.