 DOI: https://doi.org/10.12345/bdai.v6i5.32068

Analysis of security risks, protection and governance and legal system construction of large-scale models

Yipeng Zhao

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

Abstract

With the advancement of artificial intelligence, large models have been increasingly applied in natural language processing, image generation, and decision support systems. While these models demonstrate powerful generative capabilities and complexity, they also pose significant security risks. This paper analyzes potential risks in large models through four key dimensions: data security, model misuse, algorithmic bias, and system adversarial attacks. Proposals for governance strategies are presented, including data protection measures, content monitoring mechanisms, and corrective algorithms. From a legal perspective, the study proposes legislative frameworks, regulatory coordination mechanisms, and accountability systems to establish a secure, controllable, and compliant application environment for large models.

Keywords

large models; security risks; protection strategies; legal governance

大模型安全风险、防护治理与法治建设分析

赵一鹏

上海数字安全科技有限公司，中国·上海 200000

摘　要

随着人工智能的发展，大模型已逐渐运用到自然语言处理、图像生成和决策支持当中。大模型虽然具有较强的生成能力以及复杂性，但是往往伴随着诸多安全风险。文章主要由数据安全、模型滥用、算法偏差、系统对抗这四个方面来分析大模型存在的潜在风险，并针对这种风险提出了数据安全防护、生成内容监测、算法偏差校正等防护治理的思路。并且由法理角度出发，就大模型安全提出了相关的立法规范和完善、监管体系协调配合以及合理追责等法治建议，以期能为构建安全、可控、合规的大模型应用环境提供制度和技术支撑。

关键词

大模型；安全风险；防护；法治建设

1 引言

大模型是人工智能的重要载体，承载着人工智能发展的重任，其在产业升级、科研创新、公共管理等领域发挥着强大的信息处理、生成作用。但是因为对大规模数据训练依赖性较强，加上算法复杂、开放接口等特点，导致其在具体运用环节面临着较大程度的数据泄露、内容滥用、偏差放大以及系统安全威胁等诸多风险，将直接地或间接地导致模型性能下降，甚至还可能引发伦理道德、法律以及社会的信任危机问题。因此必须从技术防护、使用规则和法律建设等方面着手形成立体化的解决方案，保证大模型在合法、合规的基础上得到充分应用。

【作者简介】赵一鹏（1992–），男，中国湖北武汉人，本科，从事网络安全、应用安全、数据安全领域的产品研发相关的研究。

2 大模型安全风险分析

2.1 数据安全风险

大模型训练需要海量的数据作为支撑，其中数据的来源、采集以及管理都直接决定了数据的安全等级。所以，在大模型训练的数据采集环节，若没有严格的权限控制及溯源，就可能造成数据的泄露或者非法使用；对于包含有用户隐私的信息来说，如果没做脱敏处理或匿名化处理，则很有可能会引发敏感信息外泄；在数据的存储和传输环节，若没有数据加密、校验的措施，就可能会出现被攻击者截获、篡改或者是插桩恶意数据的情况，影响训练样本的质量以及模型的质量 [1]。

2.2 模型滥用与生成风险

大模型有很强的扩展性与生成能力，如果没有规范约束其使用行为，便极易被运用到不当场景当中。比如，部分使用者利用大模型在自动生成生产虚假信息、恶意代码或具有误导性的文本等内容，威胁网络安全以及给社会带来负

面的舆论影响 [2]。因为无法完全控制模型生成的结果，所以可能会被不法分子恶意操控并输出敏感信息。另外也有使用者恶意运用大模型，绕过安全防护，实施诈骗以及网络攻击，造成极为恶劣的社会影响。

2.3 算法偏差与歧视风险

大模型算法性能在很大程度上取决于算法训练数据，若数据集中出现失衡或是存在偏见，则可能使得模型存在不良倾向。在部分决策场景中，出现模型偏差问题会造成一些群体的不公平待遇，甚至会演变成针对一些群体的系统性歧视问题。如果一味追求算法结构与优化目标与效率或准确率，则可能会导致其普适性和公平性受到忽视。比如，在招聘、金融授信、司法评估等领域，假设模型产生了歧视性的结果，则会对社会公平造成极大的冲击。算法偏差带来的不仅仅是模型本身的公信力减弱的问题，还有可能会产生合规和伦理等方面的问题，将会对大模型的推广使用产生不利的影响。

2.4 系统安全与对抗攻击

大模型作为复杂的大规模计算系统，可能会受到一些对抗性攻击。攻击者可制造对抗样本，让模型输入出现细微扰动的情况下出现输出错误的风险，更有甚至还会使整体系统失效。此外，如果在应用环节，模型接口没有设置访问权限或是调用监控，就极易被反复试探和利用，导致模型参数泄漏或是出现安全漏洞。

3 防护治理路径

3.1 数据安全防护与合规控制

对于大模型存在的数据安全隐患，可以由构建完善的数据安全防护体系与合规控制机制着手实施有效防治。第一，加强数据采集源头的合规性把关，重点进行数据合法性、完整性、真实性的校验，结合差分隐私、数据脱敏及匿名化的技术手段降低数据的隐私泄漏风险。第二，对于涉及到数据信息的存储及传输环节均应采取端到端的加密措施及多维度的身份认证措施，并结合动态密钥管理来避免数据被非法窃取或篡改的风险。在数据存储和传输过程中需要搭配上数据完整性和篡改溯源的技术手段保障数据在整个使用期间处于无改动的状态。第三，对于数据处理和数据调用的过程而言应通过建立分级的访问权限控制措施将数据的访问范围与行为限定于特定范围当中，且能够通过日志追踪与实时监控机制实现可追溯性，防止出现未经授权的行为 [3]。第四，从法律法规以及行业准则相关的规定角度出发建立相应的审计、合规评价等制度，定期核查数据使用的合法性、合规性，保证数据能够在大模型运行的应用过程中保持安全流转且处于有效监管之下。

3.2 生成内容监测与使用限制

由于大模型在内容生成环节存在大量滥用的现象，因此有必要对其构建多层次的监测与约束体系，进行长效治理。首先，在输出的内容监测层面，对模型生成的结果采用实时的文本和代码分析，检测其语义的一致性，进行敏感词检测和风险点的分析，借助规则库、行为特征模型进行自动判定并予以自动阻断。其次，在使用权限的管理上，对不同的用户设置相应的分级访问、调用权限。采取身份验证、限制调用频率与审计机制等方式来降低非授权行为或是异常访问的情况，防止模型被滥用。再者，将动态内容的过滤和可控的生成机制嵌入生成接口处，对生产目标、输出内容以及类别予以限制，确保模型在安全、合规的基础上完成输出。在高危场景进行双次生成内容的把关和评判机制，利用人工审查和算法判断相结合的方式来实现精准把控，做好历史生成记录的存档工作，以便在事件发生时进行溯源分析。此外，加强模型的行为检测，通过分析调用方式、生的特点以及互动过程来识别其中可能存在的安全风险，防止模型遭到网络攻击、诈骗等威胁。最后，将动态更新的风险库、策略规则及监控模型相结合，形成闭环的管控制度体系，保证模型在不同场景下的输出能够始终保持稳定与可控，且能够随时对于潜在被滥用的问题进行及时处理 [4]。

3.3 偏差抑制与模型校正机制

针对大模型的算法偏差和歧视等问题，可以基于技术和制度等方面构建出多维度的抑制、校正措施。首先，数据预处理环节。对于算法样本可能存在的一些不均衡或者偏差性的问题进行检查，量化评估训练集中的群体代表性，采取欠采样、过采样以及合成数据等方式来实现数据分布的优化，避免出现不平衡的风险。其次，在大模型训练的过程中，引入相关的公平性约束和正则化目标，让算法优化不只是考虑算法本身的优化效果，还兼顾算法公平性以及稳健性的要求，避免因为过分关注预测准确程度而导致群体差异的加剧。再者，在评估和验证模型的阶段，还可建设多维度指标的检测框架，不但要关注其准确性，还要注重公平性、鲁棒性及不同人群子集的表现，将可能存在的偏差及时辨识并予以修正。此外，采取后处理的方式来矫正结果，可选用如重新分配的方法、阈值修正、概率再加权等方式，减少出现歧视结果的概率。不仅如此，还可建立健全大模型的运行测试系统，不断完善动态监控与反馈改正的机制，便于发现偏差并及时作出修改。

3.4 系统防护与对抗鲁棒性提升

针对大模型应用存在的系统脆弱性问题，必须采取多层次的防护手段。一是加强接口管理，在接口层设置好严格限制访问和控制调用次数的规则，并利用实时流量检测以及异常检测的方法来抑制恶意用户通过频繁的试探来获取模型或者模型中敏感参数信息的现象。二是从模型的输入端入手，向系统中引入对抗训练和输入随机化的手段，可降低在对抗样本附近进行小扰动带来的威胁。三是对于分布式训练部署结构而言，需要进行大量的参数运算，若发生节点被入侵或者发生参数篡改，则会使该模型本身的性能遭到破坏，

因此在分布式训练与部署架构环节也需要有加密通信协议以及参数完整性验证机制等保障网络传输的安全可靠。四是采用可信计算以及隔离执行环境技术，以多层面保护训练和推理过程，防止攻击者利用漏洞破坏系统的重要环节。五是建设多模态检测的动态防御体系，充分融合日志追踪、异常行为分析与应急响应机制，以便于在出风险初期就能将其识别并及时定位与处理，实现风险的有效阻断。

4 大模型安全法治建设路径

4.1 立法完善与制度供给

大模型安全治理需要从立法层面给予基本支持。现有法律大多从数据安全、网络安全、个人信息保护等方面作出相应规定，对于大模型范围规制还不够精准，对于大模型特殊风险也缺少针对性的规定。因此需要通过专门立法或者对有关法律条文作出修改补充来明确责任边界，规范操作方法。一是制定详细的法律条款，明确规定大模型的训练数据合规性、内容生成安全性及跨境传输行为合法性，构建起清晰的合规框架。二是应当按照行业特点完善配套的实施细则，厘清模型提供方、使用方以及第三方平台三方的权责，并对相关方提出必要的合规审查和备案要求。三是应当通过推进标准化建设，建立数据处理技术、算法透明度、输出受控等行业规范和技术标准，实现立法与技术的对接，从而通过标准为大模型安全保驾护航。

4.2 监管体系的多层协同

大模型的安全监管涵盖了社会、产业以及技术等诸多领域，单个部门无法做到综合性的立体化监管，因此要多维度协同、多部门合作地综合监管。第一，需要从国家层面上设计好总体规划并建设统筹协调机制，明确具体监管职责，形成统一的政策导向。第二，在行业层面可以设立专业监管部门或者行业自律部门，对于大模型的各个具体应用场景做出细化监管，并且进行安全评估和风险预警工作。第三，在地方层面也应该根据地域特色建立相关的联动机制，在技术审查方面加强数据安全管理和应用的合规审查，以此来实现上下联动和跨部门之间的协同治理。第四，还需建立起社会监督和公众参与的机制，运用披露信息、接受检举、独立审查等多种方式，使得多元的主体参与到监管中去，以此扩大监管范围并增强监管效果。

4.3 责任认定与法律救济

大模型一旦出错即可能带来严重的损失，如何认定责任主体、采取何种救济途径将会对法治体系执行力以及权威性产生巨大影响，所以必须要予以充分重视。一方面要通过立法的方式来明确划分模型开发者、提供者与使用者的责任，并对各环节的责任范畴以及过错性质予以区分。如若是数据泄露或者内容被滥用、算法歧视的情况，需要根据过失的主观性以及具体因果关系来追究相应责任。一方面构建起良好的法律救济渠道，提供行政申诉、民事诉讼以及仲裁调解等多元化的救济方式来帮助由于运用大模型而遭受损害的组织或个人。另一方面，积极建设跨境争议解决机制，以合理协调与救济国际上运用大模型而出现的法律冲突问题，从而从容应对大模型全球化应用带来的法律挑战。

5 结语

大模型的安全治理是技术创新与法治建设并重的系统性工作。在通过风险识别的前提下，使用数据安全防护、生成内容检测以及算法偏差校正等技术手段能够降低大模型存在的潜在风险隐患。并且完善立法、健全多层级监管和厘清归责规则是实现大模型安全的技术支撑，通过合理应用上述技术手段以及相关法律法规，二者相辅相成才能真正意义上保证大模型的安全。今后还要随着技术的发展不断改进监管的方法，从而使人工智能大模型安全可控。

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] 本刊编辑部.聚焦大模型 决胜数字时代[J].软件和信息服务（原：软件世界）, 2023, 000(12):2.

[3] 王笑尘,张坤,张鹏.多视角看大模型安全及实践[J].计算机研究与发展, 2024, 61(5):1104.

[4] 宋时磊,杨逸云.大语言模型的主权,安全及其治理[J].中国高校社会科学, 2023(6):109-118.