

生成式人工智能数据获取风险与规制路径

柴懿庭

天津大学法学院, 天津

收稿日期: 2024年6月17日; 录用日期: 2024年8月13日; 发布日期: 2024年8月22日

摘要

生成式人工智能技术的应用对于社会发展和技术进步具有重大作用, 但与此同时, 由于技术限制、缺少具体应用规范等产生诸多风险。在生成式人工智能技术的运行中, 数据获取是其基础也是风险源头, 因数据来源多样, 收集数据缺少明确的授权容易产生违规风险; 不实信息、价值观偏差、文化单一等产生收集数据质量风险; 海量数据传输过程中产生数据泄漏风险。为了防范、应对生成式人工智能数据收集引发的数据风险, 应当确立从细化人工智能立法、到明确数据收集阶段各参与主体的职责、再到强化行政监管的风险规制路径。

关键词

生成式人工智能, 数据收集, 风险规制

Generative AI Data Collection Risks and Regulatory Paths

Yiting Chai

Law School, Tianjin University, Tianjin

Received: Jun. 17th, 2024; accepted: Aug. 13th, 2024; published: Aug. 22nd, 2024

Abstract

The application of Generative Artificial Intelligence (GAI) technology plays a significant role in the development of society and technological progress, but at the same time, it generates many risks due to technological limitations and lack of specific application specifications. In the operation of generative artificial intelligence technology, data collection is the basis and source of risk, due to the variety of data sources, the lack of clear authorization to collect data is prone to violation of the risk; inaccurate information, value bias, cultural singularity and so on, resulting in the quality of the collected data risk; data leakage risk in the process of massive data transmission. In order

to prevent and cope with the data risks caused by generative AI data collection, a risk regulation path should be established from refining AI legislation, to clarifying the responsibilities of each participant in the data collection stage, to strengthening administrative supervision.

Keywords

Generative Artificial Intelligence, Data Collection, Risk Regulation

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

2022年, OpenAI 旗下一款人工智能技术驱动的自然语言处理工具 GhatGPT 横空出世, 标志着人工智能技术从“决策式人工智能”到“生成式人工智能”的重大进步[1]。决策式人工智能是目标导向性 AI, 根据已经定义好的规则、逻辑, 解决用户提出的特定问题; 生成式人工智能是创造性 AI, 能够通过人机交互不断的训练、学习, 从而生成具有创造性的新内容。生成式人工智能对于未来发展的作用在于, 可以通过改进预测、优化运营和资源配置, 为个人和组织提供个性化的数字解决方案、为公司提供关键性竞争优势、支持有益于社会和环境的科研成果。但在生成式人工智能应用过程中, 由于当前整体技术水平有限、具体应用缺乏限制与监管等原因容易致生风险, 对个人利益甚至公共利益造成损害, 因此有必要发现潜在风险源并提出规制方案。

2. 数据收集是生成式人工智能运行中的首要风险源

数据与算法是生成式人工智能的核心[2]。数据, 是“科学实验、检验、统计等所获得的和用于科学研究、技术设计、查证、决策等的数值”, 算法, 是“对特定问题求解步骤的一种描述, 是指令的有限序列。其中每一条指令表示一个或多个操作”[3]。生成式人工智能的算法逻辑比普通人工智能更具先进性和复杂性, 而算法设计、模型的先进性依赖于海量、高质量、多样化的数据资料, 但目前算法模型的开发与数据管控并非同步并行, 而是算法技术升级整体优先于数据管控。因此, 技术开发过程中往往忽略应用数据存在的问题, 不可避免因为数据瑕疵问题为应用数据安全埋下隐患。

根据生成式人工智能的内部运行原理, 除算法开发外, 其在生成文字、图片等最终成果前会经历三个阶段: 一是数据收集阶段, 即抓取大量数据形成数据库, 投入投算法域的基础阶段; 二是通过算法进行数据处理与训练阶段, 即通过算法与数据的交互, 使得算法模型具备利用数据理解和生成新内容的能力, 并在交互中不断进行优化。生成式人工智能的数据训练量巨大, 并随着每次技术迭代不断增长, 但数据量训练需求与数据质量不成正比, 数据质量无法满足日益飙升的数据需求, 根据相关研究, 高质量语言数据将在 10 余年内耗尽[4]; 三是数据输出阶段, 通过用户输入指令、算法在现有数据库中提取必要数据、数据重新组, 进而输出成果。各阶段中采用的数据集按照来源分类可分为作数据、外采数据、标注数据和人机交互数据[5]。合作数据, 是指通过合作协议调用数据接口传输、获取合作方的数据资料, 此种模式下的数据获取难度较低, 但存在偶发性与合作方之间的授权风险; 外采数据, 是指通过网络爬虫技术, 将网络空间内的所有字段返回本地存储, 形成自有数据集, 通过此种技术获取数据资料的难度极高, 需要精细的技术支持, 且存在采集数据错误、数据回溯过程中发生阻断等多种技术问题[6]。标注数据, 是指经过人工选择进行标注的数据, 此类数据获取难度低, 但人为因素较重、偏好明显, 且具有明

显的倾向性立场。人机交互数据，是指在生成式人工智能运行过程中通过人机交互对话形成的数据集，此种数据获取难度低，但存在信息失实的问题。

由于数据来源复杂多样、难度各异，导致是生成式人工智能技术在数据收集阶段引发风险的概率较高，所以数据收集既是生成式人工智能运行的基础，也是风险源头。

3. 数据收集阶段的风险划分

3.1. 开发者存在违规获取数据信息风险

以 ChatGPT 为例，生成式人工智能的运行基础在于海量数据信息形成的信息库，除本身内部资源外，还需要利用大量自然语言输入来训练自然语言模型，从而构建自然语言语料库，以生成适用于人类阅读的标准答案文本。在信息库构建和更新上，存在主动构建和被动构建两种形式。被动构建，是指用户在使用 ChatGPT 过程中，通过与其对话形式，输入个人身份、事件等信息，相关数据将被保存在数据库中。主动构建，指的是利用爬虫技术自动收集互联网上的大量数据，这种数据收集方式的优势在于可以快速高效地获取大量数据，从而获得更加准确可靠的模型训练和性能。但无论是主动还是被动收集的数据，均可因为没有获得收集及使用的实体收到，存在非法获取风险[7]。因为在数据获取权利依据不足，意大利政府公开宣布封禁 ChatGPT 有运行[8]，加拿大隐私专员办公室也公开宣布调查生成式人工智能数据与个人信息安全问题[9]。

3.2. 生成内容存在数据质量风险

3.2.1. 数据信息失实、虚假风险

随着自媒体时代的发展，网络空间每时每秒都在上载海量信息数据，但信息上载并不具备任何核查机制，用户需要通过自己的价值观判断信息真伪。如曾因媒体发布虚假最高法司法解释，导致法官在审判案件中引用该司法解释做出判决的“乌龙审判”事件[10]；流媒体平台为吸引人眼球，经常出现类似“本月起，离婚不再需要对方同意”，“头孢配酒不会致命”的标题及内容。当生成式人工智能抓取相关数据后，其生成的内容就是基于虚假数据产生的内容，从而生成误导信息、甚至在敏感领域对终端用户造成无可挽回的损害。不再具有真实性和可信性，辨别其真伪将更加困难。

3.2.2. 价值观偏差风险

第一，数据多元化缺失。目前生成式人工智能无论是 ChatGPT 还是 Stable Diffusion，多选用的是境外来源数据库。如某博主通过指令要求绘出“一人家包饺子”的画面，但生成图片画面主要元素为“欧美人”包“包子”，需要更加详细的指令“亚洲人”“中国人”定义“饺子”后，才能生成更偏重于中国风的画面。单一的数据库导致了文化包容性和多样性的不足，存在严重的文化和价值观偏差。CPT-3 的训练数据集中，中文语料仅占 0.1%，在目前世界上通用的 50 亿大模型数据训练集中，中文语料占比也只有 1.4%。数据集代表性不足容易引发统计和计算偏差，进而产生系统性偏差。

第二，为了算法模型生成内容更好的展现出人类表达，训练数据过程中为了提升算法模型性能，通常需要人工指令、监督、调整等机制。但这一过程中需要使用大量的人类偏好标签，即使有标准的规范和操作流程，但仍然存在因为反馈人员文化水平不一、道德素养不同、综合素质参差不齐，从而不经意地将其价值观差异、隐藏偏见等输入模型，从而影响候选答案排序的公正性和普遍性[11]，甚至生成毒害内容。

3.2.3. 数据时效性导致的过期风险

数据的时效性直接影响内容的可信期限，算法抓取某临界时间点的数据后，其知识范围也将局限在

该时间节点，已经训练的内容再加入新出现内容需要更加先进的技术，生成内容存在局限性和边界性。目前，OpenAI 并未给自有数据集设置专门的机制核验事实的准确性与真实性[12]，且数据更新时间节点尚不清楚。数据过于庞大，个人或单位无力承担大规模时效核实工作，加剧原有信息失实的风险。

3.3. 用户存在数据安全风险

3.3.1. 侵犯个人信息风险

生成式人工智能在对信息数据的获取过程中，尤其对个人数据的挖掘侧重于挖掘信息的广度和深度，如未来生成式人工智能商业拓展方向之一是满足特定场景下的特殊需求，在为用户打造定制化人工智能时，会首先收集用户的工人信息、了解个人偏好甚至家庭、职业、财富等敏感信息。由于算法技术对于数据深入分析的限度和边界不清，导致本不应被获取的个人信息受到侵犯，从而加剧信任危机。

3.3.2. 数据信息泄漏风险

生成式人工智能在对信息数据的过程中，在算法和数据的交互过程中，数据泄漏的风险极高。大部分生成式人工智能系统基于升级需求，需要不停的累积数据库及训练，也通常在隐私协议中要求用户授予处理个人数据及派生数据的权利。根据生成式人工智能的部署特性，在完成用户各类需求后，即会收集到海量数据，并将它们运用到后续的任务中。结合科技公司的资源和渠道，用户和算法之间形成了有效的“数据飞轮”模式[13]。当收集的庞大数据在系统中流动和传输时，即存在泄漏风险。

4. 数据收集阶段的风险规制路径

为了防范、应对生成式人工智能数据收集引发的数据风险，应当确立从细化人工智能立法、到明确数据收集阶段各参与主体的职责、再到强化行政监管的风险规制路径。

4.1. 细化人工智能领域立法

《生成式人工智能服务管理暂行办法》已经正式施行，体现了我国面对生成式人工智能带来的法律挑战时的责任与担当。但相关法律条款仍是从较为宏观的方面体现出规制的方向和原则，无微观配套的实施细则。且人工智能引发的风险更多呈无序状态，很难在单一场景下进行统一的规制。2024年1月19日，欧盟委员会、欧洲议会和欧盟理事会完成了《人工智能法》的定稿，对于人工智能回获取、应用数据、监管等提出了更加明确的要求，对于人工智能及数字经济发展具有历史性意义。如(44)用于训练、验证和测试的高质量数据集需要实施适当的数据治理和管理实践。用于培训、验证和测试的数据集(包括标签)应具有相关性和足够的代表性，并在最大程度上不存在错误，而且从系统的预期目的来看应是完整的。(45)欧盟委员会建立的欧洲共同数据空间，以及促进企业间和政府间在公共利益方面的数据共享，将有助于为人工智能系统的培训、验证和测试提供可信、负责和非歧视性的高质量数据访问[14]。为了更好的应对人工智能带来的挑战并与国际接轨，可以在《生成式人工智能服务管理暂行办法》的基础上，参考《人工智能法》的具体要求，细化实践中的操作办法。

4.2. 明确数据收集阶段各参与方的责任义务

对于生成式人工智能的开发者来说，其应当采取必要措施保障数据质量和数据安全，确保数据来源可靠、文化多元、符合社会主义核心价值观，定期进行风险检测、开展数据合规、审计等；保障数据透明，进行清晰简洁的说明和解释；配合监管机构的监督。对于提供商来说，及时检测、测试数据生成效果，确保生成式人工智能系统平台的数据安全，避免发生数据泄漏等；确保数据使用目的合法合规，保留交易记录，建立数据安全管理制度等通过数据合规保障数据安全、防止数据泄漏。

4.2.1. 利用数据合规防范数据风险

数据合规是互联网企业常用的合作模式，是指为了避免海量数据丢失、泄漏、损坏、滥用等确立的一系列标准和规定，互联网企业遵循相关标准进行数据的收集、存储、管理[15]。数据合规为数据风险防控提供了较为完备的范例，开发者/提供者可以根据《生成式人工智能》《数据》《个人信息保护法》等相关法律法规的要求，建立内部合规团队、细化内部数据安全准则、强化合规要点、定期测评监督内部流程是否符合合规要求、组织内部数据合规培训等，及时发现风险点。

4.2.2. 强化个体信息掌控和自决能力

由于生成式人工智能逻辑架构复杂，普通用户很难全面、准确认知其背后运行逻辑和数据提取机理，无法准确认知其授权背后的风险，使得用户无法作出与其真实意思表示一致的安全的合理的决策，从而引发个人信息泄漏危机。所以，为了强化用户对于个体信息的掌控和自决能力，应当面向生成式人工智能开发者构建以用户为中心的数据透明义务体系[16]。虽然生成式人工智能的算法模型逻辑复杂，但开发者仍然可以以简洁、易懂的方式为用户解释数据处理机制并提供合理指引。如全面告知输入个人数据后的影响和后果；也可通过生成式人工智能使用过程的模拟衡量解释价值。此外，还可以使用弹出式警告对相关条款进行警示，以确保用户以有限的技术知识了解系统的运行逻辑，以防止在可信度不足的情况下产生失真或有偏见的输出。

4.3. 明确行政机关对于生成式人工智能数据应用的监管职能

4.3.1. 明确行政机关的监管智能

《生成式人工智能服务管理暂行办法》第4条第1款规定，人工智能生成的内容应当体现社会主义核心价值观，不得含有颠覆国家政权、推翻社会主义制度，煽动分裂国家、破坏国家统一，宣扬恐怖主义、极端主义，宣扬民族仇恨、民族歧视，暴力、淫秽色情信息，虚假信息，以及可能扰乱经济秩序和社会秩序的内容[17]。该条规定体现了国家对于生成式人工智能内容的价值观导向。所有人工智能技术及其算法开发与更新中，都蕴含了其自身的价值观。既然算法需要依托于数据运行，在进入中国市场后，即应根据《数据出境安全评估办法》等法规，构建人工智能开放的协同体系中的中国数据出入境评估体系[18]。公权力机关应当明确监管责任、监管范围、落实监管责任分配制度，同时做好政策评估以及公司数据合作的评估。

4.3.2. 强化提供者或开发者面向监管的透明义务

虽然明确了生成式人工智能开发过程中需要履行的安全评估、算法备案、执行变更和注销备案的程序以及履行数据披露义务，但数据透明及披露义务的履行程序尚未明确，且未对其实施主体进行限定，如何提起、程序如何、披露限度如何等诸多问题尚待细化。因此，基于这一局限性，我国的治理机制应考虑强化披露义务，适当扩大提供者数据披露义务的对象类型，将对监管方的数据透明义务延伸至使用方。

4.3.3. 构建灵活高效的数据监管工具体系

美国、加拿大构建了系统化的算法影响评估制度，将治理节点前置，以准确、动态地捕捉算法技术和应用可能带来的风险。我国也应参考相关模式，构建切实可行的算法安全评估制度以及数据应用评估制度，针对数据获取过程中可能发生的风险展开系统评估。而且，因为生成式人工智能技术具有多维属性，在不同领域应用中需要获取不同的数据信息，以针对性的完成特殊场景下的需求。所以针对生成式人工智能的监管体系工具需要更加科学、高效。

4.3.4. 探索人工智能监管沙盒模式

监管沙盒是源于英国监管局的一种创新监管理念，是指设立一个受监督的监管测试区，同时匹配限

制性条件和风险管理措施,允许测试区内的企业在真实的市场环境中,以真实的用户为对象进行新技术、服务、商品测试[19]。通过测试区,有助于减少创新理念进入市场的时间与潜在成本,并降低监管的不确定性。现阶段,我国金融科技、汽车安全监管沙盒试点工作[20],对于具有创新性但风险系数高的科技来说,监管沙盒能在容错率高的前提条件下估计创新,从而避免新技术不配批准或批准后造成不可控风险的局面。就生成式人工智能而言,可以在公共卫生领域以及面向初创企业开辟监管沙盒,结合生成式人工智能的数据风险特性构建一系列监管标准,以较低的成本进行监管试验。同时,由监管机构组织设立专家组,根据标准识别并应对潜在风险。

4.3.5. 拓展算法备案制度在数据风险规制中的效能

我国《互联网信息服务算法推荐管理规定》首次提出了算法备案制度,该制度不仅有利于监管机关提前研判算法风险,还可以基于备案的算法构建合理的风险应对模式,提升算法透明度。但目前的算法备案主要针对算法主体、产品、功能几个部分备案,可以针对生成式人工智能算法与数据的交互模式提出针对性的备案要求,要求开发者出具备案报告,就数据权利、数据质量、数据安全等方面进行综合评估,以便于监管机构掌握算法对于数据处理的模式,从而判断、应对、监控可能引发的风险。

5. 结语

生成式人工智能在应用中可能引发现实世界中的社会问题,如数据安全危机、错误信息收集及生成、侵犯个人隐私、偏见和歧视等,传统治理模式难以应对新兴技术的底层、固有风险,因此需要更适当的、更具包容性的法律治理框架对生成式人工智能的开发与应用进行合理规制。而这一过程,也需要社会各界的广泛参与,坚守科研伦理,确保技术创新始终不突破道德底线,为智能生态系统提供法律保障,从而维护用户数据主体权利,实现生成式人工智能的包容性发展。

参考文献

- [1] 郭小东. 生成式人工智能的风险及其包容性法律治理[J]. 北京理工大学学报(社会科学版), 2023, 25(6): 93-105.
- [2] 商建刚. 生成式人工智能风险治理元规则研究[J]. 东方法学, 2023(3): 4-17.
- [3] 刘城霞, 段瑞雪. 数据结构与算法: Java 版[M]. 北京: 北京理工大学出版社, 2022: 6-7.
- [4] 康骁. 行政法如何应对生成式人工智能——基于算法、训练数据和内容的考察[J]. 云南社会科学, 2024, 4(4): 80-88.
- [5] 高泽晋. 潘多拉的魔盒: 人工智能训练数据的来源、使用与治理——面向 100 位 AI 开发者的扎根研究[J]. 新闻记者, 2022(1): 86-96.
- [6] 禹卫华. 生成式人工智能数据原生风险与媒介体系性规范[J]. 中国出版, 2023(10): 10-16.
- [7] 刘艳红. 生成式人工智能的三大安全风险及法律规制——以 ChatGPT 为例[J]. 东方法学, 2023(4): 39-43.
- [8] Carlini, A., et al. (2023) Italy's ChatGTP Ban Attracts EU Privacy Regulators. <https://www.reuters.com/technology/germany-principle-could-block-chat-gpt-if-needed-data-protection-chief-2023-04-03/>
- [9] OPC Launches Investigate (2023) Office of the Privacy Commissioner of Canada. https://www.priv.gc.ca/en/opc-news/news-and-announcements/2023/an_230404/
- [10] 王选辉, 马玉萱. 多地法院曾引“最高法意见”判案, 最高法裁定书: 没出台过[EB/OL]. https://www.thepaper.cn/newsDetail_forward_8505409, 2020-07-31.
- [11] 孟令宇. 从算法偏见到算法歧视: 算法歧视的责任问题探究[J]. 东北师范大学(社会科学版), 2022, 01: 1-9.
- [12] 毕文轩. 生成式人工智能的风险规制困境及其化解: 以 ChatGPT 的规制为视角[J]. 比较法研究, 2023(3): 155-172.
- [13] 张凌寒. 生成式人工智能的法律定位与分层治理[J]. 现代法学, 2023, 45(4): 126-141.
- [14] The European Union (2024) Artificial Intelligence Act.

<https://www.ey.com/content/dam/ey-unified-site/ey-com/en-gl/services/ai/documents/ey-eu-ai-act-political-agreement-overview-february-2024.pdf>

- [15] 毛逸潇. 数据保护合规体系研究[J]. 国家检察官学院学报, 2022, 30(2): 84-100.
- [16] 郭春镇. 生成式 AI 的融贯性法律治理: 以生成式预训练模型(GPT)为例[J]. 现代法学, 2023, 45(3): 88-107.
- [17] 张欣. 生成式人工智能的算法治理挑战与治理型监管[J]. 现代法学, 2023, 45(3): 108-123.
- [18] 姚志伟, 李卓霖. 生成式人工智能内容风险的法律规制[J]. 西安交通大学学报(社会科学版), 2023, 43(5): 147-160.
- [19] The Information Commissioner's Office (2023) Sandbox Assessment Criteria Indicators
<https://ico.org.uk/media/for-organisations/documents/2618128/sandbox-criteria-indicators.pdf>
- [20] 市场监管总局, 工业和信息化部, 交通运输部, 应急部, 海关总署. 关于试行汽车安全沙盒监管制度的通告[EB/OL]. https://www.gov.cn/zhengce/zhengceku/2022-04/02/content_5683112.htm, 2022-02-25.