doi:10.3969/j.issn.2097-3853.2025.05.010

生成式人工智能数据来源的合法性探究

刘霞,方宝论

(福建理工大学 法学院·知识产权学院,福建 福州 350118)

摘要:近年来以 ChatGPT 为代表的生成式人工智能技术迅猛发展,与此同时其数据来源也隐含着个人信息和知识产权侵权以及获取方式不正当等合法性风险。在目前已有相关侵权案例发生的背景下,可根据生成式人工智能数据来源侵权的实际情况,适用个人信息保护制度和知识产权法律法规,将相关法律规范与原则嵌入生成式人工智能数据来源领域,以规范生成式人工智能数据来源的获取、使用等行为。在此基础上,明确相关主体责任,完善数据收集、处理、使用等技术,协同企业、行业以及政府部门多方构建相应的监管模式,进一步规范人工智能数据来源的合法性,可以更好地解决人工智能新技术带来的挑战。

关键词:生成式人工智能;数据来源;合理使用;避风港规则;全链条监管模式

中图分类号:D913　　**文献标志码**:A　　**文章编号**:2097-3853(2025)05-0482-06

Research on legitimacy of data sources of generative artificial intelligence

LIU Xia, FANG Baolun

(School of Law and Intellectual Property, Fujian University of Technology, Fuzhou 350118, China)

Abstract: In recent years, the generative artificial intelligence technology represented by ChatGPT has developed rapidly. At the same time, its data source also implies legal risks such as infringement of personal information and intellectual property rights and improper acquisition methods. Against the background of existing relevant infringement cases, according to the actual situation of data source infringement of generative artificial intelligence, personal information protection system and intellectual property laws and regulations can be applied, and relevant legal norms and principles can be embedded in the field of data source of generative artificial intelligence to regulate the acquisition and use of data source of generative artificial intelligence. On this basis, the responsibilities of relevant subjects can be clarified, and the technologies of data collection, processing and use can be improved, the legitimacy of artificial intelligence data sources can be standardized by constructing corresponding supervision modes within enterprises, between industries and government departments, so as to better solve the challenges brought by new artificial intelligence technologies.

Keywords: generative artificial intelligence; data source; reasonable use; safe harbor rules; whole-chain supervision mode

随着人工智能技术的迅速发展,数据作为一种重要资源对生成式人工智能(AIGC)训练模型构建和深度学习起着至关重要的作用。而数据来源的合法性直接关系到人工智能系统的可靠性和公正性。因此,在判断数据来源的合法性时,要注意数据本身的权益不被侵犯以及收集、处理数据的方式要合法、正当。以 ChatGPT 为代表的生成式人工智能具有广泛的自主学习和深度学习能力,能够获得海量的信息并通过神经网络模型将其进行整合生成,实现和用户之间的高完整度交互。然而,生成式人工智能海量的数据来源面临着数据隐私、版权侵权和获取方式不当等合法性

收稿日期:2024-06-03

第一作者简介:刘霞(1982—),女,安徽安庆人,讲师,硕士,研究方向:知识产权。

问题。学界已有研究指出,生成式人工智能在数据领域具有数据源合规性风险、算法风险以及数据泄露风险三类风险,涉及算法异化和歧视、个人信息、商业秘密和国家秘密等方面[1];GhatGPT 产品在数据来源方面存在个人信息泄露风险、侵犯知识产权问题和数据抓取行为的竞争风险[2]。数据来源的合法性和合规性是生成式人工智能应用中需要考虑的重要问题。

生成式人工智能数据来源的类型众多,从内容和权利两方面对数据来源进行分类,以相关案例和研究为切入口,能够以全新的角度对生成式人工智能数据来源的合法性问题进行探讨。在此基础上,通过相关个人信息保护制度的运用和已有知识产权法律法规的适用,践行法律规范,引用法律制度与原则,将其嵌入到生成式人工智能数据来源,可有助于降低生成式人工智能的合法性风险。另外,目前关于生成式人工智能数据来源方面的侵权责任认定体系还不健全,需要进一步明确侵权责任认定,构建全链条分级分类的监管模式和制定针对性行业政策,改进数据收集技术、提升数据质量,以更好地应对人工智能新技术带来的挑战。

一、生成式人工智能数据来源的现状和合法性典型案例

当前,生成式人工智能数据来源类型各式各样,主要指的是创建新颖内容所依赖的大规模数据集。这些数据集包含文本、图片、声音、视频等多种类型的信息,用于训练机器学习模型,使其能够生成新的原创内容。促成生成式人工智能爆火的一个原因是其输入数据进行机器学习的模型取得了突破。以 2017 年谷歌团队提出的 Transformer 预训练语言模型为例[3],它最早应用于大规模自然语言场景的处理,并综合了模型中存在的现有数据实现了与用户的高度交互。生成式人工智能数据来源便是构建这一训练模型的基础本源。

(一)生成式人工智能数据来源的类型

1.公开数据集和非公开数据集

从内容来看,生成式人工智能的数据来源可分为公开数据集和非公开数据集。公开数据集是指由不同的企业、组织、政府或个人公开发布的数据,其中包括学术界和开源社区共享的数据集。利用爬虫技术来获取这类开放平台的数据可实现以较低的成本和较少的时间获取大量的数据信息,但技术手段必须合规。另外,公开数据集会表现出不成体系、较为庞杂的特点,涉及个人隐私信息以及违法数据信息等,作为数据来源也会具有合规风险。

非公开数据集包括仅供特定组织或个体访问和使用的私有数据集以及个人的数据信息等,这些数据通常不会被公开或共享,包含了大量的敏感信息。由于非公开数据集包含大量的真实数据和敏感信息,因此可以用于训练更加准确和可靠的机器学习模型。但是,在使用该数据集时,也需要遵守相关法律法规和伦理规范,确保数据的合法性和道德性。

2.享有知识产权的数据和其他权利的数据

从权利来看,生成式人工智能的数据来源可分为享有知识产权的数据和含有其他权利的数据。享有知识产权的数据是指用于生成式人工智能训练学习的数据本身享有知识产权保护,如享有著作权的作品、著作、文献等,对于这类数据的使用首先要获得授权许可。含有其他权利的数据是指除享有知识产权保护的数据之外的受到其他权利保护的数据,如姓名权、肖像权等人格利益的数据,这类数据虽然没有知识产权保护,但是会涉及权利人的其他相关权益,在将其作为生成式人工智能数据来源时需要注意是否会对他人或社会造成侵害。

(二)生成式人工智能数据来源的合法性典型案例

1.何某诉上海某 AI 科技公司网络侵权责任案

本案中,被告开发运营涉案软件以实现与用户的聊天互动。该软件运用人工智能技术及计算机算法实现整体系统的人工智能化,并经由用户输入不同形式的语料(包括文字、图片、声音等)对 AI 角色进行“调教”。在系统运行过程中使用语料筛选、分类、存储等技术,对语料进行处理和加工,并产生了相应的个性化输出。该涉案软件在未经原告许可的情况下,出现了能够指向原告的姓名、图像组合等信息,并且允许用户创设、“调教”包含自然人姓名、肖像的 AI 虚拟角色,从而吸引用户使用。被告的行为属于对原告姓名、肖像的营利性使用,其行为侵犯了原告的姓名权、肖像权和一般人格权,最终判令被告赔礼道歉、赔偿损失。

2.上海新创华公司诉某 AI 公司动漫形象版权侵权责任案

该案被告经营着一个名为 Tab（化名）的网站,该网站具有人工智能对话和人工智能生成绘画功能。被告未经授权复制了本案涉及的《奥特曼》作品,侵犯了新创华公司的复制权。此外,本案涉及的一些生成图片在保留“迪迦奥特曼复合体”作品原始表达的同时形成了新的特征。被告的行为构成了对本案涉及的《奥特曼》作品的改编,侵犯了新创华公司的《奥特曼》改编权。

基于案例 1 可知,若生成式人工智能在未经权利人授权时使用其姓名、图像等个人信息存在侵犯他人人身权利的风险。从案例 2 可知,生成式人工智能在获取数据时若未得到相应授权许可,则会侵犯他人知识产权,尤其是对版权作品的侵权。法院审理此类案件时,需要权衡人工智能技术的发展与版权保护之间的平衡。

二、生成式人工智能数据来源面临的合法性挑战

当前,生成式人工智能在数据收集和使用方面存在多个问题,主要包括侵犯个人信息权益、侵犯享有知识产权的数据、获取训练数据的方式可能不合法以及来源数据本身可能存在内容不合法或虚假错误等问题。

（一）侵犯个人信息权益

我国《生成式人工智能服务管理暂行办法》（以下简称《暂行办法》）第 11 条规定了“提供者对使用者的输入信息和使用记录应当依法履行保护义务,不得收集非必要个人信息”,这表明了人工智能相关企业在收集涉及个人信息数据来源时应当坚持必要性原则。同时,个人信息保护法规定,企业在收集个人信息时要遵守“限于实现处理目的的最小范围”的原则。生成式人工智能产品在收集数据时同样也要遵循最低限度以及合理使用的范围。然而,在实践中生成式人工智能要遵循知情合规的要求去获取海量信息以构建其训练模型是难以实现的。人工智能公司可能会通过“在合理的范围内处理个人自行公开或者其他已经合法公开的个人信息”或兜底条款去证明其收集和使用个人信息的合法性,但往往存在收集范围与限度不明的问题[4],甚至包含大量未经授权的个人信息,这将会给个人隐私安全带来隐患。

此外,以 ChatGPT 为代表的生成式人工智能进行深度学习所依赖的语料库的数据信息包含着特定组织或个体访问和使用的私有数据,这些数据通常不会被公开或共享。同时,用户在使用过程中留存的私密、敏感的个人信息、商业秘密甚至公开渠道的个人信息数据,都可能涉及主体切身利益。当生成式人工智能模型将这些数据信息纳入自身语料库并留存在神经网络后,便会产生数据泄露的风险。这些信息泄露或者被黑客盗取都可能给用户带来很大的损失,如上述案例 1,如果生成式人工智能未经授权许可而随意获取用户留存的可指向他人姓名、图像等数据进行训练时,势必会侵犯他人人格、隐私等权益,从而造成该产品因侵权无法进入市场或无法良性发展下去。

（二）侵犯享有知识产权的数据

生成式人工智能数据来源还包括一些具有知识产权的数据来源,如享有著作权的期刊文献、歌曲以及图像等。若未经提前授权而直接作为训练数据的来源,便会造成侵权问题,现实中这些信息难以逐一获得授权。人工智能公司在未经版权主体授权的情况下,对互联网上受版权保护的内容进行大量抓取并且直接收入数据库进行加工修改和二次创作,则很有可能侵犯他人作品的复制权等相关权益。此前,对于 OpenAI 在没有付费的情况下使用维基百科、WebText 的文章来训练 ChatGPT,已有国外新闻媒体对其行为进行了指责,认为其行为构成法律上的侵权。[5]可知,生成式人工智能在进行数据模型训练过程中易存在引用他人作品数据造成侵犯他人著作权的风险。而随着人工智能技术的发展,科技公司与版权单位的矛盾也日益突显,若不进行合理规制则容易导致大规模的知识产权剽窃。

（三）获取训练数据的方式不合法

通常人工智能会先把作为创作素材的作品数字化处理转换成标准数据格式,再进行深度学习。[6]这一过程主要表现为三条路径:在互联网上直接抓取已经数字化的数据作品,把非数字格式作品转化为机器可读的数字格式,将已经数字化但格式不兼容的数据进行标准格式的转换。这三种方式都是对原有内容的复制和再现,并且会在机器中形成永久的复制件,属于著作权法上的“复制”行为,存在侵犯复制权的风险。[7]现在大数据学习大部分是通过爬虫技术来获取数据,对

这些数据及内容进行挖掘使用。如果上述数据本身存在权利壁垒,那么通过上述方式所构建的训练数据模型则天然具有著作权侵权风险。

(四)来源数据本身内容的合法性问题

由于生成式人工智能的数据获取过程欠缺人工监督和价值判断,因此很难对相关信息进行实质性筛选和过滤,数据本身可能存在内容不合法不合规、内容虚假或错误等问题,从而导致其在真实性、客观性、准确性与科学性上存在不足的原生风险。另外,部分数据内容还会涉及算法歧视、偏见或侮辱、暴力等不良导向,具有突破研发者对其设置的道德伦理及法律底线的危险。同时,预学习阶段对数据库信息无良好的过滤机制,输出的内容真假难辨,增加了用户核对信息真伪的成本。因此,数据内容本身存在的"缺陷"问题会造成其作为数据来源的合法性问题。

三、生成式人工智能数据来源的规范路径

对于当前生成式人工智能数据来源面临的合法性挑战,亟待提出相关的规范路径。首先,可通过参照个人信息保护制度,确保数据收集遵循最小必要原则,保障用户知情权和选择权。其次,可适用现有知识产权法律,合理扩展使用原则和避风港规则。最后,优化责任认定,建立严格责任制度,加强监管,提升数据收集技术,以保障数据质量和安全。

(一)个人信息保护制度的参照适用

在个人信息保护法等相关法律法规对个人信息利益作出了保护性规定的前提下,行业部门在探究生成式人工智能数据来源合法性规范时,可参照使用这些法律法规,着力将其具体要求嵌入到生成式人工智能数据来源的技术运行过程中,同时监管主体对此要进行严格的监督以促进技术合规发展,并为收集训练数据过程中涉及的个人数据信息合法性问题提供规范依据,达成良好的事前合规效果。

1.个人数据信息收集遵循最小必要原则

《暂行办法》第 11 条和个人信息保护法第 6 条都规定了处理个人信息的最小必要原则。因此,人工智能企业在收集涉及个人信息的数据作为训练数据来源时,应当遵循这一原则确定合规的收集范围。最小必要原则的本质是比例原则在个人信息领域的应用,主要是为了平衡个人信息保护与新兴技术发展之间的矛盾,在新兴技术挖掘个人信息潜在价值时避免公民合法权益受到侵害。因此,在实践中要以此为出发点建立起个人利益与人工智能技术发展之间合理的利益平衡机制,如设置人工智能收集个人信息的限制条款,规定个人信息内容可供收集使用的部分以及禁止收集使用的敏感信息等事项,从而明确数据收集的范围和限度,避免侵犯他人合法权益。

2.确保信息主体的知情权和选择权

人工智能企业在收集有关个人信息的训练数据时,还需明确告知用户收集、使用个人信息的具体范围与目的以及数据可能存在的泄露风险,确保用户知情同意权和选择权的实现。2023 年正式施行的《互联网信息服务深度合成管理规定》就有明确规定:生成式人工智能产品在采集包含用户图像、声音等识别信息的训练数据时,对于涉及的个人信息隐私要以经过用户知晓和同意为前提,并在相关情形下,充分给予使用者拒绝提供的权利,如在人机交互过程中,涉及收集用户的信息时应予以显著标识进行说明,给予用户选择拒绝的权利。此外,还应建立"退出机制",即在用户个人信息数据被收入人工智能语料库后享有退出的权利,如设置数据用户专门的邮件通道,让用户对其尚未公开的数据拥有退出数据收集的选择权,从而使得信息利益主体在相关人格权利上获得最大的保护和自由。

(二)现有知识产权法律法规的适用

1.合理使用原则的扩展适用

我国著作权法第 24 条规定了"个人学习、欣赏""适当引用""科学研究"等合理使用情形。可将这一合理使用原则扩展适用到生成式人工智能数据来源方面,如爬虫技术的使用,特别是有关公开的网络信息内容,只要其属于通过搜索爬虫、在遵守爬虫协议的前提下获取的数据,则均可视为来源合法的数据。实际上,爬虫技术无法识别被抓取内容的著作权问题,如同搜索引擎一样,其抓取时也无法逐一获得版权授权。因此,可根据数据来源引用的具体情景判定善意侵权的责任。但这里存在一个问题,即权衡合理使用和人工智能商业性用途之间的关系。目前生成式人工智能的商业性应用与合理使用的本质"非商用目的"存在冲突,数据训练模型的数据能否扩展适用版权保护中的"合

理使用”原则,这是影响生成式人工智能技术发展的核心问题。拆解和妥善解决这个问题不仅要考虑著作权法的相关规定,同时也要再审视生成式人工智能内容生成的技术逻辑。[8]

2.避风港规则的类比适用

避风港规则主要应用于互联网侵权领域,适用于网络服务提供者。由于生成式人工智能新技术的出现类似于早期的互联网诞生,那么在生成式人工智能训练数据来源方面,也可以对避风港规则进行类比适用,并在此基础上建立起合规体系。

首先,对于生成式人工智能平台或网络服务提供者来说,他们可利用技术手段对其收集的来源数据进行初步过滤,尽量实现其使用的训练数据来源合法。如果接到版权人关于训练数据侵权的通知,平台应立即停止使用相关数据,并采取必要措施防止进一步侵权。

其次,生成式人工智能平台还应加强对训练数据的版权审核和管理,对此企业可建立专门的知识产权管理部门,发挥专业特长,尽力检索数据的权利状态,避免使用未经授权的数据。这可以通过与版权人建立合作关系、使用开源数据或获取授权的数据集等方式实现。

最后,生成式人工智能平台应建立有效的侵权投诉处理机制,以便版权人或其他利害关系人能够便捷地提交侵权通知。平台在收到通知后,在立即停止使用相关数据基础上迅速进行核实,最终在确认侵权的情况下及时删除或修改相关训练数据。

此外需要注意的是,避风港规则并非是绝对免责条款,同时也应当关注红旗原则在本领域的体现。即使人工智能开发商以及相关网络服务提供者在接到侵权通知后及时删除或屏蔽了违法数据来源,但如果其明知或应知存在侵权行为,或者从侵权行为中直接获得经济利益,那么他们仍可能承担相应的法律责任。

(三)责任认定方面的优化

1.明确侵权责任的认定

明确数据来源的合法性和合规性是侵权责任认定的基础,由于生成式人工智能技术的复杂性和创新性,目前在法律上还没有形成一套完整、统一的侵权责任认定标准。即使获取数据来源方式正当合法,但如果在使用这些数据内容时没有获得必要的授权或许可,也可能构成侵权。因此,在使用数据之前,必须确保已经获得了相关权利人的授权许可以及数据的使用符合相关法律规定。

对数据的处理和使用方式也是侵权责任认定的重要因素。在认定侵权责任时,还需要考虑过错责任原则。如果数据提供方、使用方或处理方在提供、使用或处理数据时存在过错,如未尽到合理的审查义务、未采取必要的安全措施等,那么他们应对由此产生的侵权问题承担相应的责任。另外,还需要考虑因果关系问题,需要证明侵权行为与损害结果之间存在因果关系,即侵权行为是导致损害结果发生的直接原因。

2.建立严格的责任制度

生成式人工智能系统的研发人员、所有者、运营者和用户在其对人工智能支配力所及范围内各自承担着责任,立法应明确其责任承担机制。在行政立法领域,国家互联网信息办公室等机构已经出台部分部门规章,对人工智能不得传播非国家规定范围内的信息作了初步监管规定,但未明确违反规定的责任主体、责任承担方式等,需要进一步明确生成式人工智能数据来源的责任主体,包括数据收集者、提供者、处理者、使用者等。这些主体应承担保障数据来源合法合规的义务,确保数据不侵犯他人版权、隐私等权益。

研发机构的程序开发决定了模型数据获取和输出方式,因此,研发机构需公开信息来源,在产生争议时证明其程序设计未侵犯他人隐私或商业秘密的故意,否则应为其“算法黑箱”承担不利法律后果。用户对他人指定为隐私或机密的非公开信息应尽合理注意义务,如故意或过失泄露数据,或以非法目的引诱类 ChatGPT 系统窃取他人隐私或商业秘密,则由该过错用户承担责任。

同时要规定合理的惩处措施。科技创新要坚持善意,不能恶意侵犯他人权利,对于恶意侵犯他人权利的行为要严格惩处并责令民事赔偿,若构成侵犯公民个人信息罪、侵犯著作权罪的,则应追究其刑事责任。

(四)明确数据来源监管和提升收集技术

1.建立全链条监管模式

2023年7月,国家互联网信息办公室等七部门联合发布《暂行办法》,提供了数据来源的合规性指引、人工智能企业的隐私政策,建立事前合规、授权的合法性基础,强调事中积极管理、自律,事后拥有完整的应急预案、监管措施,以达到全过

程动态治理的目的。基于此,相关主管部门应构建和完善对人工智能系统的事前、事中、事后全链条监管模式。

首先,可要求企业内部设立专门的审查机构,并要求技术部门配合设置专门审查软件初筛,再进行人工筛选,对数据采集内容进行合规审查。其次,训练数据包含个人信息的收集、使用应严格遵守最小必要原则,行业和政府监管部门要将常态化审查和不定时抽查相结合,建立黑白名单,精准、有效地进行监管和规制。最后,建立投诉、举报机制,通过用户举报的形式加强社会监督,从而保证已投入市场的人工智能系统不偏离法治轨道。[9]

2.明确分级分类监管体系

《暂行办法》第 3、第 6、第 16 条明确了对生成式人工智能服务进行分类分级监管的原则,强调必须有序推动公共数据的分类和分级开放,扩大高质量的公共训练数据资源,并指出由国家有关主管部门根据生成式人工智能服务适用的不同领域制定相应的分类分级监管规则或指引,进行行业部门监管。《中华人民共和国数据安全法》第 21 条也体现了数据分类分级的监管思路,通过行业主管部门参与制定监管政策,使人工智能监管法规更具针对性,由立法强化生成式人工智能的监管体系。目前我国并未提出非常明确的分类分级依据以及操作办法,可以考虑由各行业主管部门根据行业特点或需求,依据生成式人工智能应用场景,制定分类分级规范指引,对应用中可能出现的风险及影响从高到低进行排序,再根据不同风险等级划定不同的监管方式,在强调技术治理与法律规范的基础上,厘清义务边界。

3.提升数据收集技术

生成式人工智能的数据来源是其能够生成高质量内容的基础和前提,在进行数据收集时应加强技术改进,研究如何控制生成式人工智能数据的质量和偏见,使其筛选出合法、优质的训练数据,以此提升数据质量,减少数据偏见,确保数据的公正性和可靠性。例如,通过“爬虫”“自动识别算法”技术获取开放平台的数据时,应添加违法信息筛查机制,遵守爬虫协议的规定。同时,利用技术手段对数据进行加密、脱敏等处理,防止数据泄露和滥用。

四、结束语

当前,生成式人工智能技术取得重大发展,在技术迭代更新过程中其数据来源的合法性问题不容忽视。对此我们应该采取多元化的解决方式,探索既合法合规又有利于科技创新的规范路径,在采取包容审慎监管的立场上,正确适用有关法律法规和行业技术规范,在鼓励科技创新的同时,平衡其可能带来的负面影响,着力建立起事前自律守正、合法授权,事中积极管理、自律,事后拥有完整应急预案、监管措施、公正处罚的全链条规范模式。

参考文献:

[1] 孙祁. 规范生成式人工智能产品提供者的法律问题研究[J]. 政治与法律,2023(7):162-176.
[2] 聂童. ChatGPT 生成式 AI 的法律风险及合规[J]. 互联网天地,2023(3):29-33.
[3] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. December 4 - 9,2017,Long Beach,California,USA. ACM,2017:6000-6010.
[4] 陈禹衡. 生成式人工智能中个人信息保护的全流程合规体系构建[J]. 华东政法大学学报,2024,27(2):37-51.
[5] 李若一,王林,贾骥业. ChatGPT 背后的知识产权风险[N]. 中国青年报,2023-02-21(6).
[6] 玛丽特·阿瓦德,拉胡尔·肯纳. 高效机器学习:理论、算法及实践[M]. 北京:机械工业出版社,2017:1-6.
[7] 焦和平. 人工智能创作中数据获取与利用的著作权风险及化解路径[J]. 当代法学,2022,36(4):128-140.
[8] 宋华健. 论生成式人工智能的法律风险与治理路径[J]. 北京理工大学学报(社会科学版),2024,26(3):134-143.
[9] 邓建鹏,朱怿成. ChatGPT 模型的法律风险及应对之策[J]. 新疆师范大学学报(哲学社会科学版),2023,44(5):91-101,2.

(责任编辑:王圆圆)