人工智能训练数据合规性探析

邹海阳　毕梦婷　浦继尧　赵　露　鄢　龙
(西南民族大学,四川 成都 610041)

摘　要:在当今数字化时代,人工智能技术的快速发展为社会带来了巨大的变革和机遇。然而,随着人工智能应用的广泛普及,人工智能训练数据的合规性问题日益受到关注。人工智能模型的训练离不开大量的数据,而这些数据的获取、处理和使用往往涉及诸多方面的考量。在这样的背景下,探讨人工智能训练数据的合规性问题,不仅是确保人工智能技术可持续发展的关键,也是维护个人权利和社会公正的重要举措。文章将对人工智能训练数据的合规性进行探讨,分析现有问题及挑战,提出相关解决方案和建议,旨在为人工智能技术的健康发展和社会的可持续进步提供参考和借鉴。

关键词:人工智能;训练数据;合规性
中图分类号:F2　　文献标识码:A　　doi:10.19311/j.cnki.1672-3198.2024.19.011

1　AI 技术底层逻辑

AI 大模型是当前 AI 技术发展的重要领域之一,不同于以往仅能进行分类、预测或实现特定功能的模型,生成式人工智能大模型(Large Generative AI Models,LGAIMs)经过训练可生成新的文本、图像或音频等内容,且具有强大的涌现特性和泛化能力。

其中文生文工具 ChatGPT 是基于 Transformer 的语言模型,Transformer 架构能够应用于自然语言处理(NLP)。以 GPT-3(Generative Pre-trained Transformer 3)为例,其拥有超过 1750 亿个参数,仅需很少的输入就能生成高度逼真和复杂的文本。因此,Transformer 模型的出现彻底改变了 AI 生成,并引发了大规模训练的可能性。

文生图工具 DELL-E 则是基于 CLIP 的语言模型,CLIP 是 Contrastive Language-Image Pre-Training 的缩写,是由 OpenAI 在 2021 年发布的一种预训练模型。CLIP 旨在将文本和图像结合起来进行预训练,从而让模型具备理解图像和文本之间的关系的能力。它的训练数据包括来自互联网的大量图像和文本,通过对图像和文本之间的关系进行学习,使得模型能够理解自然语言描述并生成相应的图像。

文生视频工具 Sora 是一个扩散模型,同时采用了 Tranformer 架构。这种架构能够将随机噪声逐渐转化为有意义的图像或视频内容。Sora 模型通过训练,学会了理解和处理文本提示,将用户的描述转化为视频内容。具体来说,Sora 模型首先接受用户的文本描述作为输入,然后利用扩散型变换器生成一系列潜在表示(latent representations),这些潜在表示逐渐接近于真实的视频数据。在这个过程中,Sora 模型通过不断地迭代和优化,逐渐生成出与文本描述相符合的视频内容。

总之,无论是 Sora 还是 ChatGPT、DELL-E3 等生成式 AI 都是基于大模型技术研发改进而来,它们本身只是模型而没有数据,因此生成式 AI 天然地要求有大量文本、图像和视频数据的“投喂”训练。在经过大量数据训练之后,用户只需输入少量文本,AI 就可以快速生成符合要求的文本、图像和视频。

2　训练数据侵权挑战

由前文所述,AI 训练所需的大量数据(包括文本、图片和视频)是基于大模型技术的天然需求,其具有一定的正当性。但在 AI 训练过程中,也出现了对他人著作权(包括文字、图片和视频)的侵权可能,由此也带来了一定的挑战。

2.1　训练数据不可控

生成式人工智能数据收集和语料库构建高度依赖数据爬虫,其训练数据除了人为建立数据库对人工智能进行“投喂”外,人工智能还可以利用数据爬虫自动在网络上爬取数据来供自己训练。对于训练数据我国《生成式人工智能服务管理暂行办法》第七条规定,生成式人工智能预训练,优化训练的数据需满足一系列合法性要求,包括来源合法性,不得侵犯知识产权,个人信息权益等。但生成式人工智能爬取的数据在范围、数量、质量等都是不可控的,其可能突破网站经营者设置的保护措施,爬取具有知识产权保护的作品,造成对著作权人的侵权。此外,生成式人工智能还可能爬取到他人的个人信息和商业秘密,造成很严重的侵权。这些自动爬取都是依赖于具有高度自主学习技术的“算法黑箱”,其行为很难控制。

2.2　训练数据缺乏透明度

基于 AI 训练过程的复杂性、技术性和未知性,普通

基金项目:西南民族大学大学生创新创业训练国家级计划项目“人工智能训练数据合规性探析”(202310656031)。

民众和相关部门很难深入了解 AI 公司的训练数据来源和使用情况，也无法知晓哪部作品以何种方式被使用。其次，现阶段，人工智能采用“算法黑箱技术”，其使用的数据内容并未公开，同时人工智能生成的内容是向特定的使用者提供的，本身并不具有直接公开性，即使人工智能使用了受著作权保护的作品，著作权人也难以发现自己的原创内容可能被大模型训练使用，而且随着人工智能的不断更新，其生成物独创性越来越高，仅凭生成内容人们无法判断出其内容是由自己作品经训练后产出的。这给监管部门在执法过程中带来了困难，也给著作权人维权带来了挑战。

2.3 训练数据侵权难以举证

近期，美国媒体《纽约时报》将 OpenAI 和微软公司诉至法院，指控二者未经授权使用《纽约时报》数以万计文章训练 ChatGPT 等人工智能。这一争端引发了公众对于大模型训练数据版权的关注，同时也反映出大模型数据侵权认定存在的难点。目前我国的法律在举证责任方面一般遵循“谁主张谁举证”的规则，著作权人需自己寻找证据证明人工智能训练数据侵犯了著作权人的利益，而该举证过程基于以下因素往往是困难的。

首先，人工智能训练数据来源不明确。许多训练数据可能来源于多个渠道，其中可能包括版权保护的内容，但数据的具体来源往往并不清晰。在这种情况下，确定侵权责任人及其行为成为一项极具挑战性的任务。其次，数据转化难以追踪。在人工智能训练过程中，原始数据经过多次转化、处理和组合，最终形成用于模型训练的数据集。这一过程中的数据流动路径复杂，难以追踪特定数据的来源和使用方式，进而增加了侵权举证的难度。最后，证据不完整。著作权人即使发现了侵权行为，为获取足够的证据来支持起诉也是一项艰巨的任务。许多数据可能被多次重复使用，其中的原始数据可能已经难以追踪，使得著作权人的举证过程变得异常困难。

3 训练数据侵权规制路径

3.1 训练数据不应纳入“合理使用”范围

在新时代大数据背景下，共享经济是现在的主流，越来越多的人在主张“个人主义让步于集体主义”。作为新崛起的生成式人工智能，它的快速发展可以极大地促进人类的进步，便利人们的生活。在这样的背景下，为了推动人工智能进一步地发展，一些学者提出将为了训练人工智能而使用现有作品进行训练的行为纳入“合理使用”范围，牺牲著作权人的部分财产权（如，复制权），让人工智能可以免费使用现有的作品进行训练，使得训练数据合规。这样的主张确实可以促进人工智能的发展，但是笔者认为该观点过分强调“让步”，未充分考虑人类著作权人的利益。首先，人工智能确实可以促进人类社会进步，但是现阶段，人工智能主要是 AI 公司用来获利的工具。AI 公司运用大量的数据进行训练，使得人工智能不断完善，生成物质量越来越高，进而吸引更多的客户使用人工智能来进行创作，让自己获取更多的利益，其本质上并不是为了“公共利益”或“集体利益”，而是为了“商业利益”。在这样的情况下，产生利益冲突的就是 AI 公司和著作权人，相对于 AI 公司而言，著作权人处于弱势地位，此时还要让著作权人作出让步明显不合理。其次，对于著作权人而言，他们的作品是个人花费了大量时间和精力完成的，作品本身就具有很高的价值。如果将其免费作为数据提供给人工智能进行训练，则会打击人类创作者的积极性，导致人类创作者的创作减少，反而违背了《著作权法》“鼓励创作”的初衷。因此，把“为了训练人工智能而使用现有作品进行训练”的行为归为合理使用并不合理。

3.2 训练数据不应完全遵循“用必授权”原则

对于人工智能自动抓取著作权保护的作品进行训练造成侵权的问题，有学者认为人工智能对于训练数据的使用应遵循“用必授权”的原则，即人工智能公司只要使用受著作权保护的作品，就需要得到著作权人许可并向其支付合理费用。但是笔者认为该观点依旧不合理。目前人工智能主要使用“算法黑箱”，基于其数据不可知、不可控的特点，著作权人依旧无法知道自己的作品是否被使用，甚至连人工智能开发者和公司也难以知晓训练数据库中哪些资料是受著作权保护的；就算人工智能公司知道其未经授权使用了他人作品，但出于利益和成本的考虑，其也有可能会选择不告知著作权人。这不仅不能让著作权人的权利得到实现，还会使得该制度“形同虚设”，不能解决实际问题。其次，人工智能数据库中的数据非常之大，如图表 1 所示，LLaMA 已知的训练数据已达 4828.2GB，其还未包括人工智能自动爬取的数据。面对如此大的数据，如果要支付费用，并征得著作权人的同意，不仅耗费巨大，且效率低下。在该产品上市之后，其昂贵的成本费用也会分摊在每一位用户身上，不利于 AI 为各行各业赋能加速。此外，如果找不到著作权人或著作权人不同意授权，人工智能训练的数据将会大量减少，不利于人工智能自身的发展。

表 1 大语言模型数据集综合分析

大模型	维基百科	书籍	期刊	Reddit 链接	Common Crawl	其他	合计
GPT－1		4.6					4.6
GPT－2				40			40
GPT－3	11.4	21	101	50	570		753
The Pile v1	6	118	244	63	227	167	825
Megatron－11B	11.4	4.6		38	107		161
MT－NLG	6.4	118	77	63	983	127	1374
Gopher	12.5	2100	164.4		3450	4823	10550
LLaMA	83	85	92		4162.2	406	4828.2

3.3 纳入法定许可范围

对于人工智能数据问题，笔者认为不能过度保护任何一方，要找到合理的方式平衡人工智能与人类作者间的利益，使两者都能更好地发展。在解决数据侵权问题时，笔者认为应从以下几方面进行考虑。

首先，针对训练数据不公开透明的问题，应要求人工智能公司公开相关的训练数据来源。在《人工智能法案》就有相应的规定要求人工智能模型的提供者应发布关于用于训练的内容（数据）的足够详细的摘要。我国的《生成式人工智能管理暂行条例》虽然没有直接规定人工智能提供者应对训练数据进行公开，但在行业部门的监管责任中提到有关主管部门依据职责对生成式人工智能服务开展监督检查，提供者应当依法予以配合，按要求对训练数据来源、规模、类型、标注规则、算法机制机理等予以说明，并提供必要的技术、数据等支持和协助。基于人工智能“算法黑箱”技术，外界难以知晓具体训练数据，但是对于人工智能开发者而言，大部分训练数据是可以溯源的。在解决数据侵权的问题上，笔者认为可以借鉴欧盟《人工智能法案》的相关规定，在我国现有的法律基础上要求人工智能公司对相关训练数据公开。一方面可以加强对人工智能公司的监管，使其提起对内部合规性的重视，减少侵权行为的发生；另一方面通过透明训练数据，可以极大地保护著作权人的合法利益，减轻著作权人维权的难度。这能很好地解决基于训练数据缺乏透明度带来的挑战，降低了著作权人进行侵权举证的困难，能够较好地保护著作权人的合法权益。

其次，在数据公开的基础上，可以将“为了训练人工智能而使用受著作权保护的作品进行训练的行为”纳入法定许可范围内，AI 公司仅需支付一定的报酬，无须征得著作权人的同意就可以将其作品用作训练。据有关消息报道，OpenAI 正在与数十家出版商洽谈内容授权协议。且在去年 12 月，OpenAI 宣布与德国媒体巨头阿克塞尔·施普林格达成了“里程碑式”合作。根据协议，OpenAI 将付费使用施普林格旗下出版物的内容，施普林格将提供其媒体品牌的内容，作为 OpenAI 公司大型语言模型的训练数据。OpenAI 公司的做法正是基于双方签订的公开协议，通过支付合理报酬从相关平台获取大量高水平数据，从而将其投入大模型训练。这种做法正符合“法定许可”的法律情形。

综上所述，“法定许可”一方面让人工智能训练使用的数据合规，有效解决了数据侵权的问题；另一方面更好地保护了著作权人所享有的权利，两者的利益得到了更好的平衡。此外，在基于“法定许可”而使用训练数据时，应当排除侵害人格权等原就属于侵权的作品，在涉及个人信息的情况下，开发者必须保证在充分利用这些信息资源的同时，保护信息主体的合法权益。因此，应将训练数据纳入法定许可范围，与将训练数据纳入“合理使用”的方法相比，“法定许可”实现了著作权人与人工智能公司二者利益最大化的平衡。

4 结论

在公开数据来源的基础上，将使用受著作权法保护的作品进行训练的行为纳入“法定许可”范围内，同时加强对个人信息的保护。这样不仅降低了 AI 公司的成本，同时满足人们的需求，促进 AI 产业的发展，实现两者的平衡，加快为各行各业赋能增速，提升各行业社会生产力，为包括著作权人在内的人民群众创造更多的社会财富，进一步激发全社会创造活力，也符合共享集体主义的理念趋势，推动创造出更多更好的作品，最终形成正向循环。

参考文献

[1] 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.
[2] 斜晓东. 风险与控制：论生成式人工智能应用的个人信息保护[J]. 政法论丛，2023，(4)：59-68.
[3] Alan D. Thompson“What’s in My AI”2023. Hugo Touvton et ai. “LLaMA：Open and Efficient Foundation Language Mode-is”2023，华泰研究.
[4] 吴叶凡. “投喂”大模型如何规范授权[N]. 科技日报，2024-02-09(005).
[5] 张春春，孙瑞英. 如何走出 AIGC 的“科林格里奇困境”：全流程动态数据合规治理[J/OL]. 图书情报知识，1-12[2024-03-11]. http://kns.cnki.net/kcms/detail/42.1085.G2.20240305.1852.006.html.
[6] 李彤. 生成式人工智能技术提供者侵权免责事由的识别重整[J]. 南京社会科学，2024，(02)：86-97.
[7] 刘金瑞. 生成式人工智能大模型的新型风险与规制框架[J]. 行政法学研究，2024，(02)：17-32.