Open Journal of Legal Science 法学, 2024, 12(5), 3261-3267
Published Online May 2024 in Hans. https://www.hanspub.org/journal/ojls
https://doi.org/10.12677/ojls.2024.125463

Hans汉斯

生成式AI数据使用中的著作权侵权风险研究

周　啸

华东交通大学人文社会科学学院，江西　南昌

收稿日期：2024年3月22日；录用日期：2024年4月18日；发布日期：2024年5月31日

摘　要

生成式AI展现出了无限潜力，但同时也引发出一系列问题，无论是在模型输入阶段数据使用的侵权，还是模型输出阶段能否构成作品，这些问题都引发当前学界极大的争论。本文聚焦生成式AI输入阶段数据使用的著作权侵权问题，明确侵犯著作权的具体化表现，解构所侵犯的复制权和改编权类型，借鉴欧盟、美国等国家的经验，并提出解决侵权问题需将生成式AI的数据使用纳入法定许可范围并降低使用费用，或引入转换性使用制度对合理使用进行扩展，同时在AI发展的整个过程中呼吁国家注重监管，只有如此才能更好的寻求生成式AI未来的发展之道。

关键词

生成式AI，数据使用，复制权，转换性使用，法定许可

Research on Copyright Infringement Risks in Generative AI Data Use

Xiao Zhou

School of Humanities and Social Science, East China Jiaotong University, Nanchang Jiangxi

Received: Mar. 22nd, 2024; accepted: Apr. 18th, 2024; published: May 31st, 2024

Abstract

Generative AI has shown infinite potential, but at the same time, it has also raised a series of issues, whether it is infringement of data usage in the model input stage or whether the model output stage can constitute a work, these issues have sparked great debate in the current academic community. This article focuses on the copyright infringement issue of data usage in the input stage of generative AI, clarifies the specific manifestations of copyright infringement, deconstructs the types of replication and adaptation rights infringed, draws on the experience of countries such as the European Union and the United States, and proposes that solving the infringement problem

文章引用: 周啸. 生成式 AI 数据使用中的著作权侵权风险研究[J]. 法学, 2024, 12(5): 3261-3267.
DOI: 10.12677/ojls.2024.125463

requires incorporating the use of generative AI data into the legal licensing scope and reducing usage costs, or introducing a transformative usage system to expand fair use. At the same time, it calls on the state to pay attention to regulation throughout the entire process of AI development, in order to better seek the future development path of generative AI.

Keywords

Generative AI, Data Usage, Reproduction Rights, Convertible Use, Statutory Permission

1. 生成式 AI 的数据使用概述

(一) 基于语言大模型的数据学习逻辑

生成式 AI 是大语言模型(Large Language model, LLM)训练下的自然语言处理系统，如 OpenAI 公司发布的聊天机器人程序 ChatGPT，当提出一个问题时，ChatGPT 可通过自身预训练模型进行处理并生成相关答案。以 ChatGPT 模型为代表的生成式人工智能技术的出现，意味着当今社会已经步入了一个围绕由数据、算法、人工智能体所作出的社会和经济决策而构建起来的算法社会[1]。生成式 AI 将神经网络算法和深度学习技术相接合，能更好的学习人类语言组成和遣词造句，使得输出端的表达更加符合我们的思维习惯。从人工智能兴起的伊始，人类就不间断的研究如何使得机器更加“拟人化”，在人工智能领域的三大学派中，联结主义学派认为 AI 就是模仿人类的仿生学，造物主创造了人类，而人类复制了自己。美国国家工程院院士杰夫·霍金斯(Jeff Hawkins) [2]说：“最终我们将构建像我们一样的智能机器，这些机器将与大脑的工作方式相似。”这些观点无不暗示着生成式 AI 的运行逻辑——像人的大脑一样进行思考和运算，而事实也确是如此展开。生成式 AI 整体可分为输入端和输出端两部分。在输入端，大语言训练模型主要由数据收集、数据预处理、数据训练环节构成。最初，以数字资源为主的海量信息通过多种渠道进行收集，其信息数据大多数从互联网、电子设备、期刊杂志等多个供给元而来，大量的原始资源被纳入构建的训练数据库中。这些数据资源使用量之庞大更是令人咋舌，以 ChatGPT-4 为例，一个能够具备通过美国律师资格证的能力[3]的生成式 AI，其所用训练模型的参数集至少一千多亿，数据资源的多寡成为影响生成式 AI 最后输出内容是否精准、丰富、连贯、多样化的重要因素。接下来，在进行训练之前需要将这些数据进行预处理，对所掌握的训练数据进行标注、划分以及纳入新的示例数据集，从而使得生成式 AI 更好的接受信息。最后，通过奖励模型和强化模型进行参数训练，并微调监督学习训练出来的模型，使得输出文本的误差最小化。在上述过程中，生成式 AI 运用神经网络架构技术进行数据的学习处理，从而使得其专业领域的学习能力更强。

(二) 生成式 AI 数据使用的范围

数据、算法和算力是数字时代的核心生产要素，数据输入阶段以数据为主，机器学习阶段以算法、算力为主[4]，人工智能的学习、成长与内容生成离不开对以数据形式表现的数字化作品的获取与利用[5]。生成式 AI 所使用的训练数据有互联网、各类数据库、社交媒体、新闻、书籍等各种来源。《中华人民共和国数据安全法》第二十七条规定：“开展数据处理活动应当依照法律、法规的规定，建立健全全流程数据安全管理制度”，第三十二条规定：“任何组织、个人收集数据，应当采取合法、正当的方式，不

得窃取或者以其他非法方式获取数据”。在当前法律制度下，用非法手段获取的非法数据当然性的属于违法行为。只有合法收集的数据才能正当的所被承认、利用。生成式 AI 的发展建立在庞大的数据资源学习基础之上，而现实中，囿于现有数据资源的知识产权保护方式，其合法获取数据的成本和速度远远落后于发展的需求，法律的滞后性已成为生成式 AI 行业发展的绊脚石。一边是成式 AI 企业对庞大数据资源需求如饥似渴，另一边则是通过传统知识产权保护方式所合法获得的数据资源如涓涓细流，从而使得一些生成 AI 企业不惜铤而走险使用非法手段获取资源，并引发了著作权利人对人工智能企业不断的知识产权侵权诉讼，这无疑反应了目前传统著作权资源获取途径已无法满足生成式 AI 发展的需要且已成为急需解决的问题。在科技竞争日新月异的今天，基于大模型技术的生成式 AI 就同提供动力的瓦特蒸汽机，将会成为产业升级的重要契机[6]，各国纷纷布局 AI 产业争取下一轮科技制高地，欧美等国不断出台政策鼓励 AI 行业的发展，我国也应为生成式 AI 的发展创造一个良好的法律环境。

2. 生成式 AI 数据使用中著作权侵权类型

(一) 对作品复制权的侵犯

生成式 AI 在训练过程中使用了大量的文本、音频、图像等数据，这些数据很可能涉及他人享有的版权[7]，其中，复制权是著作权人最为重要的权利，在新的物质载体上加以固定或在新的作品中保留了原作品的基本表达都是侵犯了著作权人的复制权。数字化别人的作品是复制行为，因此在数据输入阶段对作品直接数字化的行为自然构成了侵权[8]。生成式 AI 所用到的训练数据资源来源于其构建的数据库，当对生成式 AI 所要利用的数据进行收集时，这一环节中所利用到的数据库往往会收集到大量著作权人未授权的作品，不管是通过电子数据的形式进行收集还是将纸质作品通过扫描的方式进行转化，若未得到著作权人的授权，那么必将会构成对作品复制权的侵犯。例如当前纽约时报等多家数字媒体起诉 OpenAI 人工智能侵犯著作权一案，纽约时报认为 OpenAI 对其生成式 AI 的数据训练未得到许可而使用了其数百万篇文章，要求 OpenAI 对其非法行为负责和赔偿。尽管当前此案还尚未了解，诉讼仍在进行当中，但生成式 AI 所引发的复制权侵权问题已成为时下关注的焦点。生成式 AI 就收集整理的数据进行训练学习时，对数据作品的利用到底在何种情况下才构成复制权侵权？这要结合生成式 AI 的运行原理进行分析。由于生成式 AI 的技术复杂性，其学习复制作品过程中对作品既有临时复制也有永久性复制。一方面，临时复制产生于计算机的运行之中，主流观点认为临时复制仅是一种技术现象，具有暂时性、附带性，不属于复制权控制的行为，即并没有侵犯著作权人的合法权益。若生成式 AI 学习的过程中所进行的作品复制为临时复制，这便是一种临时性的、无害的客观行为。并非所有的复制都构成侵权，而是要根据其实施后的行为来判定。但若将复制而产生的销售、展览等变现所得，则会构成著作法上的侵犯复制权。另一方面，若生成式 AI 的数据学习并非只将数据短暂复制于系统中，而是需将作品数据长时间停留，此时则会对作品产生永久性复制，即构成复制权的侵权。综上所述，生成式 AI 的数据训练学习中对复制权的侵犯风险极大，所涉及的作品类型也多种多样。

(二) 对作品改编权的侵犯风险

《著作权法》规定，改编权是改变作品创作出具有独创性新作品的权利，新作品脱胎于原作品而又独立于原作品，是改编权的外在形式特征，首先，新作品必须源于原作品，若与原作品毫无关系，根本谈不上对原作品的改编，对于一个形成的新作品来说，它应保留原作品的基本表达。其次，新作品还需具有著作权法意义上的独创性，这样才能够得上著作权意义上的新作品。生成式 AI 在数据训练学习过程中对原作品的利用程度将会成为判断其是否侵害原作品著作权人改编权的关键，在人工智能创作中，如果最终输出的生成内容虽具有一定独创性，但仍然保留了数据库中某一作品或者某些作品的基本表达，应属于改编作品[9]。在实践中，生成式 AI 是否构成对原作品改编权的侵犯还要结合具体情形来看待。

以生成式 AI 绘图为例，生成式 AI 可经过充分的训练后，根据用户给出的指令最终生成图像。在生成的过程中，AI 根据学习训练的数据作品为依据，通过数码手段将图像作品转化成为机器可以处理的代码，而此过程中其最终生成的作品是在收集的庞杂的数据库里通过分析筛选并基于大量元素的数据集合所产生，最后产出的作品各式各样，带有很强的创新型，且难以对标改编权的原作品，因其不确定性的选择和难以控制的作品输出而无法与侵犯改编权的行为产生密切联系，因此对作品的改编权几乎没有什么影响。但如果生成式 AI 训练所使用的作品具有特定性，并在输出端带有特定作品的特定特征，则将会构成对作品改编权的侵犯。如微软公司开发的 AI“下一个伦勃朗[10]”，工作人员将伦勃朗生前的作品数字化并用以训练 AI，使其在完全学习伦勃朗技术的基础上生成具有新颖性、独创性的作品，上述行为是 AI 在基于原作者作品基础上形成新的作品，其目的在于模拟、再现某个作者的作品，尽可能地与作者的写作风格相吻合，所以这样由人工智能抽取出来的信息，实质上是作者个人的个人表现，因此势必会侵犯原作品著作权人的改编权。所以，这种表现式的“机器学习”行为属于版权意义上的对作品的使用，不构成“合理使用”，而是对作品的侵权使用，应当承担版权侵权责任。

3. 生成式 AI 当前著作权法适用困境

(一) 与著作权法传统许可模式的冲突

传统的著作权许可使得创作者授权他人在一定时间、一定范围内按照约定的方式使用其作品，而生成式 AI 对于数据利用的庞大需求给传统著作权许可模式带来了巨大挑战。在生成式 AI 的数据使用中，必须逐一获得原作品版权所有人的授权并缴纳一定的许可费，才能规避侵权风险；但实际应用中，因数据量的庞大性，生成式 AI 的数据使用完全获取大量用户的一对一授权十分困难，这就使得交易双方均无法有效地利用作品。于此同时，生成式 AI 数据使用要向权利人付费，其用数据数量非常庞大且种类繁多范围宽广，由此会产生高额的许可费用，若数据使用成本的原因远超收益成本，则会使生成式 AI 企业无法接受。我国著作权法同时亦规定了法定许可和合理使用两种制度，但目前两者的明确列举式规定都难以适用生成式 AI 的数据使用情形，对于生成式 AI 的数据使用来说，法定许可制度还存在着收费难的问题，而合理使用制度也有着一些弊端。

(二) 合理使用制度的不足

我国著作权法第 24 条中对合理使用作出了明确的规定，该条款规定在十三种情况下用户可以不经版权所有人的允许以及付费，就可以对其作品进行使用。2020 年《著作权法》修改后，用“不能影响作品的正常使用，也不得不合理地损害著作权人的合法权益”取代了“不得侵犯著作权人依照本法享有的其他合法权利”，从而正式引进了三阶段测试标准，但这一修改并没有改变我国版权保护的封闭立法模式，因其规定较为缺乏灵活性，所以法院须在法定限度内对未经许可的作品加以规制，从而在面对像生成式 AI 等新兴行业问题时顿感捉襟见肘。除了列举的“合理使用”以外，著作权法还对其他情形的适用作出了相应的规定，其他情形适用合理使用制度的前提是它必须符合其它法律、法规的要求，所以，仅仅从著作权法角度来看，我国的司法机构还不能在现行的“合理使用”范畴之外再界定出新的合理使用情形。因此在没有生成式 AI 数据使用单独针对性的立法条款下，此条款的使用规定也无法适用。综合来看，在当前合理使用制度的现状下，尚无生成式 AI 数据使用这种新兴事物的适用之地，也使得我国著作权法有关合理使用制度的变革愈发重要。

4. 域外经验借鉴

(一) 欧盟：数字中介服务

2022 年欧盟理事会批准的《数据治理法》作为欧洲数字经济发展战略的一项重要内容，其核心之一

就是以数据中间商为主体，构建以信任为基础的数据流通和交易生态。数字中介服务在一定数目的数据主体、数据持有者和数据用户之间，通过技术、法律或其他方式，为进行数据分享而形成了一种商务关系，其中包括对数据主体的相关权利的行使。人工智能企业可以在数据中间商进行数据交易，且不需要征得他人的同意就可以利用其数据或材料，而数据提供者则可以获得合理的经济补偿，形式上符合我国法定许可制度的模式[11]。

(二) 美国：转换性使用规则

美国作为英美法系国家的代表，其针对合理使用制度的发展主要来源于其自身的司法适用和相关判例。美国法院运用其版权法第 107 条中的“四要素检验”法对合理使用原则进行了探讨。四项要素之间存在着一定的联系，根据对作品使用的目的与性质的评价来判定一项行为的是否为合理使用，该要素的核心问题主要在于使用行为的可转换性，这也是判定四要素的重点。转换性使用的关键是使用行为有没有添加新的事物，有没有存在更多的意图或特征，有没有以新的表达方式、意义或信息来改变原有作品，尽管有被复制的客观事实，但是复制的目标与原作的表现价值无关，而是产生与原作不同的其它内容，独立于原作品的存在，使得普通读者即便细心地阅读也不能会出与原作同样或相似的结论。

5. 生成式 AI 著作权风险的规避路径

(一) 纳入法定许可制度

法定许可是指在向作者支付一定的费用后，无需征得作者的同意就可以进行使用。在互联网这个由用户自创内容的时代，由于创作的作品种类繁多、数量庞大，一次次授权的代价太高，使得人工智能企业很难一一获得版权所有人的授权。或许在做出了合理的解释后，再去适应新技术所带来的新行为，才是最好的选择。有些学者认为，要保证高质量的工作，首先要保证作者的利益，AI 并没有制造出“新作品”的能力，如果让 AI 来剥夺创作者的权利，那就是“杀鸡取卵”了，生成式 AI 的法律规制应该平衡各方的利益，维护公共利益和社会秩序，保障个人利益和私人权益[12]。因此，在当前情境下，扩大适用法定许可的行为，将生成式 AI 纳入其中，既可提高授权的效率，并给与著作权利人一定的回报，使作者能够受到激励，创作出更好的作品，也可使生成式 AI 获取充足的数据来学习和提高算法的能力。将生成式 AI 的数据使用纳入法定许可制度，还能保障版权人的权益以及满足国家对人工智能产业发展的需求。本文认为，法定许可能够更好的被利益双方所接受，可在法定许可下，成立生成式 AI 版权授权集体管理协会，通过官方的平台进行作品登记，并向权利人支付报酬，如此集约式管理还可进行集体议价，更好的降低作品使用费，从而达到在降低交易成本下保证生成式 AI 数据学习与保护著作权人利益的平衡。

(二) 引入转换性使用发展合理使用原则

从当前我国《著作权法》中关于合理使用的规定来看，生成式 AI 的数据训练所对作品的使用无法涵盖其中。如何在激励创作与促进作品传播利用之间取得平衡，解决权利排他与表达自由的矛盾，实现推进文化艺术科学事业发展的制度宗旨，一直是著作权法的重要议题[13]。合理使用是一种对著作权的限制机制，它在被设计的时候就是为了消除著作权权利所带来的障碍，从而可以用来激励和推动创新，激励创新，推动知识的公共传播。在我国司法活动中，司法者对合理使用的认定虽然具有灵活性，但是合理使用的司法解释仍须遵循“三步检验法”之限制[14]。在著作权法中对于哪种情形下能够被纳入合理使用，从而作为受专有权利控制的例外而无需向作品权利人支付报酬或经其许可，所依据的判断标准并不是看它是否被规定于权利限制这种形式的条款中，而是看其是否符合三步检验的标准。从实质上看，生成式 AI 的作品数据使用需满足上述三步检验法，才有纳入合理使用的可能。但三步检验法较为缺乏灵活性，加之我国著作权法较为封闭式的立法，在合理使用所列举的情形之外的其他新兴事物难以被合理的规制，因此，在当前我国合理使用制度的基础上，可借鉴美国“转换性使用”标准，并将其引入合理使用制度

当中，以此用来规制生成式 AI，从而更加合理的平衡发展著作权保护的需求。美国《版权法》在 1976 年制定了四项标准，以确定合理使用：1) 使用作品的性质和目的；2) 受著作权保护的作品属性；3) 使用的数量和内容的实质性；4) 对可能的市场或作品价值的影响。美国的 Leval 法官把对“合理使用”的第一个要素(使用的目的与性质)描述为审理“合理使用案件的核心”。

目的要素是“转换性使用”四要素中的主导要素，只要“转换性使用”成立，就可以认为是“合理使用”；对原作的“转换性”程度越高，其“合理使用”的可能性就越大，即使是商业用途的使用也可以被认定为“合理使用”，这是由于“转换性使用”推动了版权立法促进文学艺术发展的最终目的。在“转换性使用”所产生的新作品中，原作对新作品的价值与作用都没有太多的贡献，所以，如果原作品权利人再向新作品创作者提出许可与付费的请求，不仅不合乎情理，且还会影响到作品的创作。生成式 AI 在使用数据训练创作新作品的过程中，所创作出的内容具有独创性表达或意义，而不是简单复制或重组文字；从目的性上来看，企业使用作品数据来训练 AI 在于使其更加智能化、拟人化，而并非简单复制使用原作品，对原作品在现实中的传播和相关受众的映射关系并无影响，因此符合合理使用的原则，也能够避免算法偏见和控制交易成本[15]。将“转换性使用”制度应用于生成式 AI 的数据学习，即可增强合理使用判断的灵活性，也将会更好的助推生成式 AI 行业的发展。

(三) 政府机关加强行业监管

生成式 AI 行业的发展方兴未艾，挑战与机遇并存，政府对人工智能行业进行及时监管，防止其野蛮生长也是其应有之义。在生成式 AI 著作权侵权问题尚无明确法律回应的当前，针对运用数据学习资料的规范以及相关侵权问题，政府监管部门和司法机关应根据当前的法律法规、行业审查机制进行合理规制。在未来，当生成式 AI 被纳入合理使用或法定许的范围后，若无其它机构监督，规模较大生成式 AI 企业很可能会出现垄断，所以，这一点亦需国家来规范，同时，在法定许可体系下，要完善对于生成式 AI 企业通过行业协会向每个作者支付的定价费用，避免因此引起市场混乱。最后，对于生成式 AI 企业的恶意侵权，应该更好发挥专侵权责任法律规范的作用。

6. 结语

新时代发展的当下，生成式 AI 的进步令全球瞩目，甚有成为下一个科技制高点之势，而我国著作权法对其规制尚不完善。本文较为浅显的提出化解生成式 AI 著作权侵权风险可通过合理使用制度或法定许可制度规制的观点，同时认为对其整个行业的发展亦离不开政府的正面引导；随着当今社会人工智能的不断发展，其未来面临的著作权侵权问题也会层出不穷，而在此过程中只有平衡好其中各方利益，处理好其中存在的法律问题，才能为人工智能行业的未来铺平道路。

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] 界面新闻. 专访美国工程院院士杰夫·霍金斯: 创建通用人工智能关键在于透析人类智能[EB/OL]. https://baijiahao.baidu.com/s?id=1769115078064070571, 2024-03-22.

[3] 侯利阳, 李兆轩. ChatGPT 学术性使用中的法律挑战与制度因应[J]. 东北师大学报(哲学社会科学版), 2023(4): 29-39.

[4] 刘少军, 聂琳峰. 人工智能生成内容的著作权法之辩[J]. 南昌大学学报(人文社会科学版), 2024, 55(1): 107-118.

[5] 张进. 论 ChatGPT 对著作权法的挑战及其应对[J]. 时代法学, 2023, 21(6): 45-57.

[6] 顾男飞. 生成式人工智能的智能涌现、风险规制与产业调控[J]. 荆楚法学, 2023(3): 70-83.

[7] 崔原, 连鹏飞, 任晓, 等. 生成式人工智能知识产权风险及应对思路[C]//中国家用电器协会. 2023 年中国家用电器技术大会论文集: 2023 年卷. 2023: 4.

[8] 赵宏伟, 茹克娅·霍加. 生成式 AI 背景下著作权侵权样态及其风险治理[J]. 网络安全与数据治理, 2023, 42(9): 59-64.

[9] 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.

[10] 李安. 机器学习作品的著作权法分析——非作品性使用、合理使用与侵权使用[J]. 电子知识产权, 2020(6): 60-70.

[11] 詹爱岚, 田一农. 生成式人工智能机器学习中的著作权风险及其化解路径[J]. 电子知识产权, 2023(11): 4-14.

[12] 曹博. 著作权法如何应对 Web3.0 挑战: 以视听内容为样本[J]. 东方法学, 2023(3): 85-97.

[13] 郑飞, 夏晨斌. 生成式人工智能的著作权困境与制度应对——以 ChatGPT 和文心一言为例[J]. 科技与法律(中英文), 2023(5): 86-96.

[14] 沈玥. 人工智能深度学习的合理使用研究[J]. 湖北经济学院学报(人文社会科学版), 2023, 20(7): 72-77.

[15] 李杨. 生成式人工智能风险的法律类型化治理[J]. 延安大学学报(社会科学版), 2024, 46(1): 31-38+57.