Dispute Settlement 争议解决, 2025, 11(8), 185-192
Published Online August 2025 in Hans. https://www.hanspub.org/journal/ds
https://doi.org/10.12677/ds.2025.118260

Hans汉斯

生成式人工智能训练数据的版权侵权风险及规制

李乔乔

南京理工大学知识产权学院，江苏 南京

收稿日期：2025年7月18日；录用日期：2025年8月8日；发布日期：2025年8月20日

摘　要

近年来生成式人工智能的爆发式发展态势给传统著作权法带来了前所未有的挑战，对于生成式人工智能训练数据的版权风险，当前著作权法中的法定许可以及合理使用制度已凸显出不适应性，因此亟待探究出对生成式人工智能训练数据侵权风险的有效规制路径。为缓解技术发展与保护文艺创作之间的尖锐矛盾，文章将立足于对生成式人工智能训练数据版权侵权成因的分析，梳理现行规制路径并反思其不足，进而提出应对策略，以期能够助力于实现科技发展与版权保护之间的利益平衡。

关键词

生成式人工智能，训练数据，版权侵权，合理使用

Copyright Infringement Risks and Regulation of Training Data for Generative Artificial Intelligence

Qiaoqiao Li

School of Intellectual Property, Nanjing University of Science and Technology, Nanjing Jiangsu

Received: Jul. 18th, 2025; accepted: Aug. 8th, 2025; published: Aug. 20th, 2025

Abstract

The explosive development of generative artificial intelligence in recent years has posed unprecedented challenges to traditional copyright law. Regarding the copyright risks associated with training data for generative artificial intelligence, the current statutory license and fair use systems in copyright law have become increasingly inadequate. Therefore, there is an urgent need to explore

文章引用: 李乔乔. 生成式人工智能训练数据的版权侵权风险及规制[J]. 争议解决, 2025, 11(8): 185-192.
DOI: 10.12677/ds.2025.118260

effective regulatory pathways to address the infringement risks of generative AI training data. To alleviate the acute conflict between technological advancement and the protection of literary and artistic creations, this article will analyze the causes of copyright infringement related to generative AI training data, sort out existing regulatory approaches, reflect on their shortcomings, and then propose response strategies. It is hoped that this can help achieve a balance of interests between technological development and copyright protection.

Keywords

Generative Artificial Intelligence, Training Data, Copyright Infringement, Fair Use

1. 问题的提出

2024 年 7 月党的二十届三中全会提出在健全因地制宜发展新质生产力体制机制的重点任务中，要求完善推动新一代信息技术、人工智能等战略性产业发展政策和治理体系，并且 2024 年 12 月上旬召开的中央经济工作会议也将开展“人工智能+”行动列为 2025 年以科技创新引领新质生产力发展的重要方向之一；同时，物联网、大数据以及云计算等前沿技术的深度融合与迅猛发展，为人工智能构筑了坚实的技术基石。在国家政策与科技浪潮的推动下，人工智能技术呈现出爆发式发展态势，ChatGPT、豆包、DeepSeek 等生成式人工智能相继问世，而生成式人工智能的崛起却给传统著作权制度带来了前所未有的颠覆性挑战。

生成式人工智能模型的不断优化需要“投喂”大量数据资源，而这些数据资源大多来自于各类版权作品，生成式人工智能产业的进步离不开海量数据支持，但对于版权人而言，他们不应成为科技产业发展的牺牲品，因此如何缓解科学技术创新与文艺作品保护之间的矛盾，让所有人共享人工智能时代红利，是新质生产力发展的重要课题。鉴于此，本文将在分析生成式人工智能训练数据版权侵权成因的基础上，梳理现行规制路径并反思其不足，进而提出应对生成式人工智能训练数据版权侵权风险的解决策略。

2. 生成式人工智能训练数据版权侵权的现状

生成式人工智能的核心逻辑是基于海量数据完成模型训练，进而通过算法自主生成全新的数据内容。可见人工智能的输入、利用以及输出阶段都离不开数据，因此分析生成式人工智能训练数据版权侵权的成因，可以从数据的收集处理以及训练输出等环节进行探究。

2.1. 数据收集与处理环节

数据是生成式人工智能模型的基石，只有经过海量的数据训练，让生成式人工智能模型经过不断学习，才能对新的数据进行分析、预测并生成高质量自然语言文本等操作，可以说数据资源是提高生成式人工智能竞争力的核心要素之一。

数据收集环节可能面临的版权侵权风险，大多在于生成式人工智能模型训练的数据来源不合法[1]。用于生成式人工智能模型训练的数据一般有以下五种来源：第一种是自有数据，人工智能开发者可以依托自身开发的其他产品或服务收集、整理数据。例如腾讯混元大模型参数规模超千亿，预训练语料超 2 万亿 tokens，这些语料数据就是腾讯的自研数据，包括腾讯旗下众多产品和服务所产生的数据，如腾讯云、

腾讯广告、腾讯游戏、腾讯金融科技、腾讯会议、腾讯文档、微信搜一搜、QQ 浏览器等 50 多个腾讯内部业务和产品的数据。自有数据权属明确，在使用过程中无需考虑版权侵权的问题。第二种是公开数据，是指通过合法渠道可以被公众获取、使用的数据，比如通过政府开放平台、学术机构数据库以及公共数据集等平台获取，通常不受版权或者隐私限制。第三种是通过购买或者交叉许可获得的数据，即通过向专门的数据提供商购买获得数据使用许可或者合作方之间交叉许可使用数据，此种方式成本较高并且需要确保数据的合法性。例如在 Thomson Reuters 诉 Ross 一案中，Ross 公司申请使用 Westlaw 数据被 TR 公司拒绝后故意向与 TR 公司有合作的 LegalEase 购买大量基于 Westlaw 案头批注编写的备忘录数据用于生成式人工智能的模型训练，此种未经授权使用受版权保护的作品进行模型训练的行为严重损害了版权人的利益，侵害了 TR 公司的版权，并且不属于合理使用的范畴。第四种是通过爬虫采集的数据，即生成式人工智能可以利用数据爬虫在互联网中自行爬取大量数据进行模型训练，此种方式能大大节约成本，但是却极有可能引发法律风险，产生严重的侵权纠纷，比如可能突破网站创建者设置的技术保护措施爬取大量版权作品，或者可能爬取到侵犯他人隐私信息或者商业秘密的数据[2]。

人工智能在获取训练数据后对数据进行挖掘分析，首先便需要将获得的数据进行复制并存储至数据库中。复制行为区分为永久性复制和临时性复制，我国《著作权法》尚未对复制行为作出明确解释，当前学术界对于临时性复制是否属于《著作权法》中的复制行为尚存争议，王迁老师认为“临时复制”只是数据分析过程中的附带步骤，并不会产生一个具有独立经济价值的复制件[3]，但也有学者认为只要构成再现作品的行为都应当属于复制，而不应局限于“固定”要件[4]，司法实践中法官对此有较大的裁量权，因此人工智能对训练数据的复制行为无论是永久性复制还是临时性复制，只要获取的数据并未经过版权人的使用许可，那么都可能构成对版权人复制权的侵犯。

2.2. 数据分析与输出环节

人工智能在储存大量数据资源后会对数据进行清洗、分析与整合，在这个过程中可能会对数据进行翻译、改编或者汇编，形成新的数据集，然后对新的数据集进行深度学习。如果新的数据集保留了原作品的部分元素，可能会有侵犯原版权作品翻译权、改编权或者汇编权的风险。但是版权人一般不会对数据分析阶段引发的侵权行为提起诉讼，因为数据分析阶段形成的数据集难以获取并作为证据固定，因此该阶段的侵权更为隐蔽。

在生成式人工智能的技术应用场景中，数据输出环节的版权侵权风险呈现出独特的复杂性。具体而言，人工智能模型输出内容与训练数据集中的受保护作品存在实质性相似甚至直接重合的可能性，构成该环节的核心风险源。由于终端用户普遍缺乏专业的版权侵权鉴别能力，难以通过直观比对识别内容来源的合法性；同时，当前数据平台对人工智能生成内容的审核机制尚不完善，大多依赖规则简单的算法过滤，审查范围局限于有限的特征匹配，无法有效覆盖语义层面的相似性检测或跨模态的数据关联分析。这种技术能力与监管需求的错配，使得数据输出环节的版权侵权行为难以被及时发现和规制，进而加剧了版权侵权风险规避的现实困境。

3. 生成式人工智能训练数据版权侵权的成因与规制困境

3.1. 生成式人工智能训练数据版权侵权的成因

3.1.1. 相关监管规定粗疏

2023 年 7 月 13 日国家互联网信息办公室联合国家发展改革委等七个部门发布规范性文件《生成式人工智能服务管理暂行办法》，该文件的出台虽然填补了我国对于生成式人工智能领域专门监管政策的空白，但是较之于人工智能领域出现的复杂且隐蔽的法律问题而言，该文件的相关规定略显粗疏。例如

该文件第 7 条对生成式人工智能服务提供者和使用者提出了要求，指出“生成式人工智能服务提供者应当依法开展训练数据处理活动，使用具有合法来源的数据和基础模型，不得侵害他人依法享有的知识产权”，该条仅仅对数据训练的合法性进行了一刀切规定，却并未对数据来源、数据清洗或者数据输出监管等具体环节进行详细规定。反观其他国家在生成式人工智能领域的法律法规，欧盟于 2024 年生效的《人工智能法案》明确，生成式人工智能需遵循欧盟《数字单一市场版权指令》中关于文本和数据挖掘的例外或限制规则。根据该规则，以文本与数据挖掘为目的，对合法获取的作品或其他内容实施复制与提取的行为，属于法定例外情形，不过版权人明确保留权利的情形除外。在文本与数据挖掘行为的法律规制方面，日本在其《著作权法》中对相关行为的合理性认定采取了更为灵活开放的态度。而美国法院在“谷歌案”中所采纳的“转换性使用”观点，则表明美国在机器学习训练数据的作品使用问题上，呈现出较为宽松的司法倾向。就我国目前对 AI 领域的治理现状而言，还需要更加完备的法律制度进行约束规制。

3.1.2. 理论规定与司法实践未达一致

目前美国已经发生多起关于使用他人版权作品进行生成式人工智能训练是否侵权的案件，多数案件尚在审理中，我国也出现了相关案例，在上海新创华诉广州年光一案中[1]，原告主张被告未经许可将奥特曼系列作品作为数据模型进行训练，并生成实质性相似的图片。法院通过独任审判仅从 AI 的输出结果对比原告提供的版权作品进行判断，认定被告侵犯了原告的复制权和改编权，而并未对 AI 数据训练过程进行评述。而在上海新创华公司诉杭州某智能科技一案中[2]，法院则认为对生成式人工智能应当持谨慎包容的态度，鼓励技术进步和商业发展。生成式人工智能的创设和发展不可以避免需要在输入端引入他人作品作为训练数据，在无证据证明生成式人工智能是为使用权利作品的独创性表达为目的以及已影响作品正常使用或者不合理损害著作权人合法利益的情形下，可以被认为是合理使用。

我国 2023 年 8 月 15 日施行的《生成式人工智能服务管理暂行办法》明确规定了用于生成式人工智能模型训练的数据来源必须合法，不得侵犯他人知识产权。可见我国当前规定对用于 AI 模型的数据要求是十分严格的，而上述案例中法院的态度则完全不同。广州互联网法院对生成式人工智能数据训练的过程不予评述，杭州互联网法院则是认为对生成式人工智能应当持审慎包容的态度，不能过度监管从而遏制了技术发展，由此可见我国理论规定与司法实践对于生成式人工智能训练数据的态度并未达成共识。

3.2. 生成式人工智能训练数据版权侵权风险的规制困境

在生成式人工智能版权侵权案件中，训练数据使用者一般会通过主张自己存在较低的注意义务、许可使用或者合理使用等抗辩理由消除其行为的违法性。注意义务是指生成 AI 开发者对其“投喂”的海量训练数据具备合理审查义务；许可使用是指使用者通过与版权人签订合同，获得在一定期限和范围内以特定方式使用其作品的权利；合理使用是指允许使用者在一定条件下不经版权人许可使用其作品。版权是利益平衡的产物，无论是侵权规则原则还是侵权抗辩事由，都在为著作权法实现技术创新与文艺保护的利益平衡中发挥了重要作用，但是，当其被运用到生成式人工智能领域时，均凸显出一定的局限性。

3.2.1. 侵权归责原则存在争议

我国当前颁布的应对生成式人工智能知识产权侵权纠纷的文件，之所以较为粗疏与空泛，是因为一般的侵权规制思路应用到生成式人工智能领域仍旧存在一定的局限性。在应对生成式人工智能训练数据的版权侵权纠纷中，侵权的具体归责原则值得商榷[5]。

[1]广州互联网法院(2024)粤 0192 民初 113 号民事判决书。
[2]浙江省杭州市中级人民法院(2024)浙 01 民终 10332 号。

若要求人工智能开发者承担严苛的注意义务，对训练数据的来源合法性实施全流程、穿透式审查，不仅需投入巨额合规成本，更会显著抬高技术研发门槛，形成阻碍创新的制度性壁垒。并且训练数据资源能够直接影响生成式人工智能模型的竞争力，若仅使用公共开发数据确实能从根本上解决侵权的风险，但当前公共开放数据的数量较少且质量良莠不齐，难以支撑生成式人工智能模型的训练，由此可见，使用海量训练数据资源时的版权侵权行为是难以避免的。反之，若放松对开发者的注意义务标准，虽能释放企业创新活力，但极易导致侵权行为的规模化扩散，使著作权保护体系面临系统性风险，动摇知识产权法律秩序的根基。同时，注意义务过低原则也并不适用于生成式人工智能训练数据的版权侵权规制。我国在互联网版权侵权治理领域引入了“红旗规则”，即只有发现显而易见的著作权侵权事实时才构成侵权，并且引入“避风港”原则，即当著作权人发出侵权通知时，网络服务商可以通过“通知－删除”行为进行免责，达到降低其注意义务的效果。“红旗原则”和“避风港”原则所针对的情形都是网络服务商的间接侵权行为，正因其间接侵权行为并无主观上的故意，因而降低其注意义务，但是在生成式人工智能领域，开发者对于训练数据的侵权使用，属于直接侵权，大多存在主观上的故意，因此这二者并不能用相同的思路进行规制。

关于训练数据侵权归责原则的争议折射出人工智能时代版权保护的深层困境——既要为技术创新预留必要的发展空间，又需维护著作权人合法权益，如何在两者间构建动态平衡机制，已成为人工智能治理领域亟待破解的重大课题。

3.2.2. 合理使用制度的局限性

根据《著作权法》的底层原理可知，合法使用他人作品的方式除了获得许可外，还包括合理使用。纵观美国、日本或者欧盟等国家，可以发现当前其政策法规都倾向于通过合理使用解决生成式人工智能训练数据合法性的问题，欧盟在《单一数字市场版权指令》中规定“以科学研究为目的”以及“不限制使用目的”条件下的数据挖掘规则，日本在其著作权法中规定了“计算机处理信息的轻微使用条款”，将商业组织或其他社会组织通过计算机技术进行数据挖掘的行为纳入合理使用制度进行规制，美国则在《数字前年版权法》中将高校科研人员以研究或教学为目的进行的生成式人工智能数据训练的行为纳入合理使用的范围，其认为基于生成式人工智能模型训练的数据使用属于“中间复制”，可纳入“转换性使用”的范围[6]。但反观我国当前合理使用制度，将其适用于生成式人工智能训练数据存在以下障碍：

首先，生成式人工智能训练数据并不属于我国《著作权法》第 24 条第一款所述具体情形。我国《著作权法》第 24 条规定了个人使用、科学研究以及适当引用等情形，个人使用主要针对个人的学习研究使用，而不能是商业主体，更不能是人工智能；科学研究少量复制需要具备非盈利的目的，但是生成式人工智能数据使用是商主体以营利为目的进行模型训练；适当引用情形是针对“为介绍评论某一作品或者说明某一问题”而不可避免地需要引用，并且引用是“适当”的，但是生成式人工智能对于训练数据是直接使用，而非“引用”，并且生成式人工智能的模型训练需要大量数据资源，也并不符合“适当”的条件。此外，第 24 条还规定了“法律、行政法规规定的其他情形”这一适用合理使用制度的兜底条款，但是我国目前除了在一些规范性文件或部门规章中对生成式人工智能的训练数据进行了规定外，还尚未在法律或者行政法规中对此问题进行规定，因而生成式人工智能挖掘、使用训练数据的行为也无法通过该兜底条款进行解释。由此可见，生成式人工智能并不属于以上合理使用的情形，因此生成式人工智能挖掘、使用训练数据的行为在我国难以通过《著作权法》第 24 条进行规制。

其次，即使对合理使用的具体类型采取开放判断模式，生成式人工智能挖掘与使用数据训练的行为也难以适用合理使用制度。一是开发生成式人工智能的最终目的是为了投入商业化使用并获取利润，再用获取的利润持续进行开发，形成正向循环，而此商业化目的将会导致生成式人工智能挖掘、使用数据

的行为无法符合“三步检验法”的标准。二是通过大量原始数据进行训练的生成式人工智能模型生成的新数据将会对原始数据的应用产生冲击，即生成式人工智能生成的作品将会替代原始作品，挤占原始作品的应用市场，将对原著作权人的利益产生损害，又悖于我国著作权法的立法初衷，因此难以适用我国合理使用制度。

3.2.3. 许可使用制度的局限性

我国版权法领域的许可使用制度主要是指法定许可、默示许可以及授权许可三种方式，其中默示许可和授权许可皆不适用于生成式人工智能数据的侵权风险规制。默示许可是指当人工智能开发者未征求版权人同意使用了其作品用于模型训练时，若著作权人保持沉默则推断该许可成立，但由于著作权人在人工智能平台“投喂”的海量数据中很难及时查询到自己的作品使用情况，因此这显然对著作权人是非常不利的。授权许可则是指生成式人工智能开发者需要经著作权人同意并向其支付费用获得将其作品用于 AI 模型训练的权利，但是模型训练需要大量的数据资源，若每条数据都需要获得授权，会大大消耗生成式人工智能开发成本，若消耗成本超过侵权成本，甚至会出现激励 AI 开发者反向侵权的现象[7]。由此可见，默示许可和授权许可对于人工智能领域版权侵权风险的规制存在较大的局限性且难以突破。

法定许可制度不仅简化了授权流程，兼顾了生成式人工智能开发者的利益，还进一步保证了著作权人的利益，尽可能实现了著作权人与生成式人工智能开发者之间的利益平衡。相较于合理使用制度和其余两种许可默示，法定许可制度优势更加明显，避免形成著作权人利益向生成式人工智能开发者的单向流动态势。虽然通过法定许可使用制度规制生成式人工智能训练数据版权侵权风险是当前最为合适的思路，但是仍旧存在许多还未解决的问题。首先，我国的法定许可制度具有较强的公共属性，设立目的是平衡著作权人利益和公共利益，保障著作权人基本权利的同时促进文化发展和知识传播。而生成式人工智能使用数据进行训练虽然从一定程度上平衡了版权人与 AI 开发者之间的利益，但是却并非为了促进知识传播等公共利益目标，并且 AI 通过模型训练后生成的内容甚至可能与原作品形成市场竞争。其次，由于生成式人工智能模型训练需要大量且不同领域的数据，因此涉及到的使用费率计算将是一项复杂且困难的工作，这将消耗著作权集体管理组织大量的人力物力并且最后是否能够完成这一挑战还是未知数，同时将海量数据的法定许可使用费交付给每一个版权人等隐形成本的消耗也是巨大的。除此之外，法定许可的适用还面临着跨越国界的困难，若是使用的数据跨越多个国家，而每个国家对于作品版权的规定都是有所区别的，那么，取得跨境作品许可的协调成本将进一步增大法定许可适用的困难程度。

综上所述，上述规制思路均存在其难以克服的局限性，导致难以适用于生成式人工智能训练数据的侵权风险应对。究其根源，是因为我国针对解决生成式人工智能问题的现有制度并不完善。

4. 生成式人工智能训练数据版权风险的规制建议

通过前述分析可以发现，当前生成式人工智能数据训练的版权风险规制主要存在两个问题：一是现有的规制思路较为粗疏，无法应对生成式人工智能领域出现的多种版权侵权风险，二是当前的法律制度存在滞后性，无法契合对生成式人工智能问题的规制。为此，应着眼于对规制思路以及法律制度的完善，在规范生成式人工智能服务提供者行为的同时也要完善合理使用制度，保护版权人权利的同时也要为人工智能稳步发展提供保障。

4.1. 形成“事前披露－事中规范－事后监管”的全链条规制

针对生成式人工智能数据训练在不同过程可能出现的多种侵权样态，应健全对涉及数据的多个阶段的规制手段，即在输入数据前应完善对数据使用的披露，在数据训练过程中应注重“非表达性使用”的目的，在输出数据时应加强对输出内容的监管。

4.1.1. 完善训练数据使用前的版权信息披露规范

数据透明度当前已成为世界各国讨论的重点议题，欧盟在《人工智能法案》中规定生成式人工智能的所有训练数据信息都应当进行披露，而美国也在《人工智能基础模型透明法案》中呈现出重视生成式人工智能训练数据版权信息披露的立法趋势，可见提升数据透明度将成为各国数据治理的重要目标[8]。2024 年 3 月，我国网络安全标准化技术委员会发布《生成式人工智能服务安全基本要求》，倡议服务提供商对语料中涉及知识产权部分的摘要信息进行披露。但是还应对具体的披露内容作出规定，即版权信息披露规范应当规定数据使用者在使用数据前对其版权信息进行披露，并规定所需披露的内容，包括数据来源以及获取渠道等。完善数据的版权信息披露规范，有助于版权人及时获取作品的被使用信息，从而更及时地进行维权。

4.1.2. 明确“非表达性使用”的训练目的

我国著作权法遵从“思想－表达二分”原则，即保护表达而不保护思想，但非表达性使用恰恰是对作品所蕴含思想的利用，而并不直接妨碍作者的原创性表达，也不会对其表达造成替代性威胁。著作权法的设立宗旨为保护原创性表达，激励创新并促进知识传播，而非是禁止不以直接使用表达为目的的复制行为，而非表达性使用即为不侵害著作权人权利的同时通过合理使用作品促进知识的传播。

但是生成式人工智能使用数据进行模型训练时如果只是对原数据的思想、情感等要素进行分析，最后生成的新内容与原数据是“不相似”的表达，因其并未生成与原作品相似的内容，也就不构成对原作品的替代，也就不构成对原作品市场份额的挤压，因此也极大减轻了对著作权人利益的损害[9]。但有学者认为若因“使用具有非表达性”而完全豁免复制行为的合法性审查，将有可能纵容对权利人复制权的滥用。关于对训练数据进行的“中间复制”行为，应当根据人工智能时代的到来对“中间复制”的含义进行重塑[10]，传统市场环境下对复制行为的定义已无法适用于当前人工智能领域，因此不能再将固定作品的行为简单粗暴定义为复制行为，为契合人工智能领域的发展，应当将固定作品并进行传播的行为定义为复制行为，因此“中间复制”行为并未经过传播便不构成著作权法中的复制，也即不构成对复制权的侵犯。

4.1.3. 强化服务者过错原则下的注意义务

关于生成式人工智能侵权在我国适用无过错责任还是过错责任一直存在争议，具体则体现在学界于 2023 年以来发布的两部人工智能法建议稿，《人工智能示范法 2.0 (专家建议稿)》则认为应当一律采取过错推定原则，而《人工智能法(学者建议稿)》则区分一般人工智能与关键人工智能，分别适用过错责任与过错推定责任[11]。但鉴于人工智能侵权并不会产生严重的损害，同时为了促进人工智能行业发展，应当对其适用过错责任原则。但是在过错责任原则之下生成式人工智能服务提供者的注意义务也应当作进一步规定。

相较于生成式人工智能服务提供者，版权人处于弱势地位，其难以及时对所拥有的作品进行监管，同时无论是用于模型训练的数据侵权还是生成的内容侵权，服务提供者都比版权人更便于也更有能力进行监管。因此人工智能平台应当发挥“守门员”作用，对人工智能输出内容进行严格审查，强化审查机制和扩大审查范围，避免陷入版权侵权风险。

4.2. 完善合理使用制度对生成式人工智能训练数据的适用

经过前述分析可以发现，合理使用制度为目前最能直接解决人工智能训练数据侵权风险的策略，首先，经过各国实践考察发现，合理使用制度是当前较为可行的路径，就我国目前人工智能发展的阶段而言，借鉴人工智能发展成熟的国家之经验更为明智；其次，数据训练为人工智能发展必不可少的阶段，

若将数据训练列为版权侵权，可能会导致人工智能产业的发展停滞，阻碍我国科技进步；而引入许可制度规制数据训练行为将会造成市场供求关系严重不平衡，并且许可制度将会消耗大量的市场成本，导致版权许可市场发生市场失灵，也将会对人工智能产业发展造成极大冲击，因此较之于许可制度，完善合理使用制度与我国当前发展现状都更为契合。

4.2.1. 增加合理使用制度的适用主体

从域外立法规定看，各国虽然对合理使用的主体范围规定不同，但是都针对进行数据挖掘行为的主体做出了相关规定，我国学界对此还尚未达成共识。随着数字经济的发展，人工智能技术已经深入到我国经济发展的多个领域，而且适用的场景仍在不断扩充。合理使用制度的设计是为了实现著作权人利益和社会公共利益的平衡，因此在尽可能不损害著作权人市场利益的情况下，可以兼顾对社会经济的考虑，增加适用合理使用制度的主体，而不仅仅局限于非商业目的的主体，从而促进人工智能技术的稳步发展[12]。

4.2.2. 扩展合理使用制度的适用情形

我国《著作权法》第24条对合理使用制度的适用情形进行了规定，虽然设置了兜底条款，即“法律、行政法规规定的其他情形”，但是仍然创设了前置条件，即需由法律或行政法规规定，导致了该条款的灵活性不足。因此为了扩展合理使用制度的适用情形，同时兼顾法律的稳定性特征，可以在《著作权法实施条例》中增设将“数据挖掘、数据训练”等行为作为合理使用的适用情形，经过实施反馈结果后再考虑是否直接对《著作权法》第24条进行修改。

5. 结语

我国人工智能技术的研发正如火如荼地进行，也因此引发了一系列风险，其中生成式人工智能训练数据的版权侵权问题便是当前争议的热点之一。生成式人工智能的技术研发需要海量的数据支撑，导致人工智能开发者在获取、使用或者输出数据的过程中将难以避免出现侵权，因此从促进技术发展的角度需要放宽对数据挖掘或训练行为的要求，将其纳入合理使用制度的适用范围。但是保障人民享受技术红利的同时不应牺牲著作权人的利益，应尽可能实现版权人利益与社会公共利益的平衡，即针对人工智能平台形成“事前披露－事中规范－事后监管”的全链条规制，以实现作品创新与技术发展的良性循环。

参考文献

[1] 陶冉. 论人工智能生成物侵犯著作权之责任认定[J]. 太原师范学院学报(社会科学版), 2025, 24(1): 7-16.
[2] 王卓群. 生成式人工智能绘画著作权侵权与规制路径研究[J]. 西部学刊, 2025(7): 75-82.
[3] 王迁, 褚楚. 人工智能与著作权边界初探: 技术进步下的法律挑战与思考[J]. 中国编辑, 2024(8): 56-62.
[4] 余俊, 孙雪静. 边缘计算中临时复制的著作权法规制[J]. 中国出版, 2022(7): 46-52.
[5] 姚秀文. 论通用人工智能训练数据版权侵权之归责原则[J]. 科技创业月刊, 2024, 37(9): 8-12.
[6] 储翔, 周怿霖. 适应生成式人工智能数据训练的版权合理使用规则完善[J]. 中国出版, 2025(6): 3-7.
[7] 金宇菲. 生成式人工智能训练数据著作权侵权风险及应对[J]. 出版与印刷, 2025(3): 38-49.
[8] 李安. 人工智能训练数据的版权信息披露: 理论基础与制度安排[J]. 比较法研究, 2024(5): 136-152.
[9] 张伟君. 论大模型训练中使用数据的著作权规制路径[J]. 东方法学, 2025(2): 79-92.
[10] 施小雪. 重塑复制权: 生成式人工智能数据训练的合法化路径[J]. 东方法学, 2024(6): 70-83.
[11] 常烨. 生成式人工智能数据“投喂”的著作权侵权行为规制[J]. 科技与法律(中英文), 2025(2): 31-41.
[12] 刘云开. 人工智能训练作品的著作权合理使用进路[J]. 东北大学学报(社会科学版), 2025, 27(1): 117-126.