Open Journal of Legal Science 法学, 2026, 14(5), 128-137
Published Online May 2026 in Hans. https://www.hanspub.org/journal/ojls
https://doi.org/10.12677/ojls.2026.145142

Hans汉斯

生成式AI训练数据的著作权侵权规制路径

赵　悦

宁波大学法学院，浙江　宁波

收稿日期：2026年3月29日；录用日期：2026年4月13日；发布日期：2026年5月21日

摘　要

生成式AI发展离不开海量高质量训练数据，其数据获取与使用常涉及受著作权保护的作品，引发产业发展与著作权保护的突出矛盾。现行授权许可、合理使用、法定许可等著作权规制路径存在实操困难、场景适配不足、机制僵化等局限，难以应对相关问题。准法定许可制度通过公告异议、默示同意、合理付酬的设计，兼顾AI开发者授权成本与著作权人合法权益，契合知识产权制度本质与国家发展战略，且有立法先例与技术支撑，具备可行性。构建该制度需明确权利义务、规范程序、完善配套，实现技术创新与权利保护的平衡，为生成式AI产业健康发展提供法治保障。

关键词

生成式AI，数据训练，著作权，准法定许可

Regulatory Approaches to Copyright Infringement in Generative AI Training Data

Yue Zhao

Law School, Ningbo University, Ningbo Zhejiang

Received: March 29, 2026; accepted: April 13, 2026; published: May 21, 2026

Abstract

The development of generative AI is inextricably linked to massive volumes of high-quality training data. However, the acquisition and utilization of such data frequently involve copyright-protected works, triggering a prominent conflict between industry development and copyright protection. Current copyright regulatory frameworks—such as authorized licensing, fair use, and statutory licensing—suffer from limitations including practical operational difficulties, insufficient adaptability to varying scenarios, and rigid mechanisms, making them inadequate for addressing these emerging challenges. A quasi-statutory licensing system, designed around mechanisms of public notice and

文章引用: 赵悦. 生成式 AI 训练数据的著作权侵权规制路径[J]. 法学, 2026, 14(5): 128-137.
DOI: 10.12677/ojls.2026.145142

objection, implied consent, and reasonable remuneration, effectively balances the licensing costs for AI developers with the legitimate rights and interests of copyright owners. This approach aligns with the fundamental principles of the intellectual property system and national development strategies. Supported by legislative precedents and technical feasibility, it represents a viable solution. Establishing this system requires clarifying rights and obligations, standardizing procedures, and perfecting supporting mechanisms to achieve a balance between technological innovation and rights protection, thereby providing robust legal safeguards for the healthy development of the generative AI industry.

Keywords

Generative AI, Data Training, Copyright, Quasi-Statutory Licensing

1. 引言

在数字时代，以生成式 AI 为核心的新一代人工智能，正于全球范围内加速兴起并得到大规模应用，其凭借强大的数据处理与学习能力，获得了理解人类自然语言的能力以及生成文本、图片、音频和视频等内容的能力[1]，这不仅深度重塑了人类知识创造与信息互动的基本模式，更在经济社会各行业、各环节中实现全方位渗透，极大地促进了新质生产力的发展。然而，生成式 AI 的数据训练需要海量数据，这其中包含着大量的作品，由于难以取得著作权人的逐一授权，极易引发著作权侵权风险。《生成式人工智能服务管理暂行办法》第七条规定，生成式人工智能服务提供者应当依法开展预训练、优化训练等数据处理活动，其中涉及知识产权的，不得侵害他人依法享有的知识产权。为了实现生成式 AI 训练数据的合法性，提升我国人工智能产业发展的可预测性，亟需根据生成式人工智能训练数据的需要，创新著作权保护与许可方式。为解决这一问题，本文在分析著作权侵权困境以及现行路径局限性的基础上，提出“准法定许可”这一制度，并论证其正当性及具体的构建思路，以期平衡生成式 AI 产业发展与著作权保护之间的关系。

2. 生成式 AI 训练数据的著作权侵权困境

自大数据储存和处理技术出现以来，人工智能开辟了数据训练算法的技术路径，其以机器学习为技术基础，作用于智力创造活动领域，其技术过程可分为数据输入、机器学习、结果输出三个环节[2]。生成式 AI 的训练数据主要来源于两个方面：一是通过网络爬虫技术对公共互联网中的海量数据进行批量抓取[3]，二是用户在使用的过程中通过提问的方式主动输入自己的信息。通常而言，训练数据的质和量决定了人工智能想象力和创作力的高低，但由于用以训练大模型的数据会大量涉及仍处于著作权保护期的作品，因而在现行著作权法框架下存在着巨大的侵权风险，进而导致了技术迭代的迫切需求与既有法律规则之间的冲突。

2.1. 生成式 AI 训练数据的海量需求

随着算法的持续演进、算力的稳步提升以及大规模数据的广泛应用，以机器学习为核心代表的人工智能技术实现了突破性发展。生成式 AI 依托数据训练库，能够从海量数据中高效挖掘符号间的关联规律

并转化为知识，将其存储于模型与数据之中，而后应用于创作场景，生成多种形式的新内容[4]。因而可以说“没有数据，就没有人工智能”。

训练数据的数量是生成式 AI 模型训练的基础前提。生成式 AI 的核心运作逻辑，在于对输入数据开展统计学维度的处理与分析。只有进一步拓展训练数据的规模层级，技术才能捕捉到小规模数据库中难以显现的逻辑关联。以面部图像生成为例，相较于在规模有限、存在偏差的面部图像数据集上训练而成的模型，依托大规模、多样化面部图像数据集训练的模型，显然更有可能生成逼真度高、形式多样的面部图像。从当前产业的发展来看，技术越先进的生成式 AI 技术，训练所需数据越多，数据规模往往越庞大。例如，GPT-2、GPT-3 的参数数量分别为约 15 亿和 1750 亿，GPT-3 在训练时更是使用了将近 45 亿字节的数据，并深度剖析了其中近 570GB 的数据[5]。

训练数据的质量影响着生成式 AI 模型训练的效果。多数生成式 AI 具有通用性特征，因而它的训练数据不能仅专精于特定领域，要尽可能的高质量、多元化，尽量使用内容全面、联系紧密且逻辑周延的训练数据[6]。以语言模型为例，其需要完成包括问答、概述、翻译等不同形式的语言任务，要求对人类语言进行高度理解并作出准确的生成，这就需要吸收来自不同领域的知识，而人类作品中蕴含着复杂的逻辑结构、独特的表达方式和深厚的知识体系，这种高质量的数据训练可以使 AI 有效捕捉到人类的需求，生成符合人类审美和认知习惯的内容。当前顶尖的生成式 AI 技术普遍需要包含作品在内的高质量数据，例如，DeepSeek 大模型先是在 14.8 万亿个 token 的高质量数据上对 DeepSeek-V3 进行预训练，然后通过监督微调和强化学习来充分发挥其能力。OpenAI 向英国政府提交的一份文件中称，“如果无法获得版权作品，我们的工具将无法运作”[7]。

2.2. 训练数据使用引发的著作权侵权风险

生成式 AI 的数据训练需要海量的高质量数据，在这个过程中，其对数据的获取和使用涉及到作品复制权等著作权的权利范围，存在较大侵权风险。具体而言，该阶段对训练数据的使用可以分为数据获取、数据输入和数据学习三个环节，各环节均面临不同的侵权风险。

在数据获取环节，人工智能模型常借助网络爬虫技术，在公共网络空间、社交媒体平台等多种渠道批量抓取数据，并将其存储至特定服务器，为后续的模型训练提供基础支撑。这些数据可能来自书籍、文章、音乐作品、影视作品等，其中包含了大量受著作权保护的原创内容。在这个过程中，如果被抓取的对象是享有版权的作品，且人工智能相关操作未获得著作权人的明确授权，那么其收集数据的行为就是对于作品的复制，会侵犯著作权人的复制权[8]。在美国司法实践中，已有多名作家联合对英伟达公司提起集体诉讼，主张该公司旗下的 NeMo AI 平台擅自利用盗版文学网站的素材，用于训练人工智能的自然语言生成能力[9]。这充分表明，数据获取环节所潜藏的著作权侵权风险，已受到社会各界的广泛关注与重视。

在数据输入环节，人工智能模型要将其前期收集的原始数据转码成为符合需求的结构化数据，在此基础上完成数据清理、分类筛选、整合优化等一系列预处理操作，最终形成与需求相对应的新数据集合，实现数据内容的针对性输入，为人工智能机器学习提供基本的数据资源。在这个过程中，潜藏着侵犯著作权人改编权、汇编权及翻译权的风险。机器学习的技术特性决定了人工智能必须将原始数据转码为标准化的结构化数据，而转码过程必然涉及对原有数据内容的多方面调整，包括数据格式的转换与修改、对冗余信息的整理删除、以及对相关数据的汇总编排等操作。这些行为会构成对著作权人的翻译权、改编权以及汇编权的侵犯[10]。

在数据学习环节，人工智能模型对转换后可解读的数据进行深度学习，剖析数据的特征属性、结构形态及相互间的关联逻辑，进而具备依据人类的指令、描述或具体要求生成对应内容的能力。训练过程

中，人工智能会不断调整自身参数以优化性能，将参数调整到最优状态，从而生成更贴合人类使用习惯与实际需求的内容。在这一反复学习、参数调整与模型优化的循环过程中，必然会涉及数据的复制行为，从而可能会侵犯作品的复制权[11]。更重要的是，生成式人工智能的产出内容会对训练数据中所涉作品形成市场替代效应，直接影响原作品的潜在市场价值。在数据训练场景下，人工智能模型通过深度学习输入作品，提炼并记忆其中蕴含的创作思路与表达风格，再以有别于原作品的形式输出内容，最终导致生成的内容与原作品既相似又有差异[12]。

2.3. 生成式 AI 产业发展与著作权保护的冲突

生成式 AI 技术的迅猛发展与传统著作权保护制度之间形成了难以调和的矛盾。这一矛盾的核心在于数据利用与权利保护之间的冲突，若不能得到妥善解决，将同时制约技术的创新与文化的繁荣发展。

若过度偏向 AI 产业发展而忽视著作权保护，将对文化发展造成严重的不利影响。在生成式 AI 训练数据的过程中，其抓取的作品绝大部分未经许可使用，会直接侵害著作权人的复制权、翻译权、汇编权等多项权利。从长远来看，还会损害创作者的利益、打击其创作积极性。一方面，生成式 AI 不同于传统技术的关键就在于它具备突破性的创作能力，能够在针对特定风格的作品进行学习后，掌握特定的表达风格，从而生成大量的具有相似特点的内容，这种模仿行为不仅可能影响作品的创作者当前的市场份额，更有可能在未来市场竞争中形成直接的对抗关系[13]。另一方面，著作权人在 AI 大规模的抓取中处于相当的被动地位，其无法得知作品被抓取，也不能控制作品被使用的情况，如果无限制地允许 AI 基于商业目的不经许可的使用作品，那么著作权人对作品享有的私权就会沦为社会公共资源，很有可能打击创作者的积极性，影响文化产业的繁荣发展。

反之，若过度强调著作权保护而限制数据获取，同样会阻碍技术创新与产业发展。作为科技革命和产业变革的核心驱动力量，人工智能已成为引领科技革命、提升国家竞争力和维护国家安全的战略技术产业[14]。然而从技术需求来看，生成式 AI 技术的发展高度依赖海量训练数据，且训练数据的规模和质量与模型能力呈现正相关趋势。严格的授权要求将使研发者面临巨大的交易成本障碍，训练数据库往往包含了海量由无数作者创作的作品，这使得在追溯相关作品的著作权人的过程中，开发者需要承担因识别权利人、开展授权谈判等交易流程产生的极高成本，而且即使联系到了著作权人，如果没有可观的报酬，其通常也缺乏授权意愿。在这种情形下，要求开发者主动联络数量庞大且分布零散的权利人进行协商并达成合意，显然不具备现实可行性[8]。

3. 生成式 AI 训练数据著作权问题解决方案的局限

生成式人工智能的蓬勃发展，使其对海量、高质量训练数据的需求变得空前迫切。然而，生成式 AI 技术的迅猛发展与传统著作权保护制度之间形成了难以调和的矛盾。现行著作权法为解决作品使用问题，主要提供了授权许可、合理使用与法定许可三种路径。然而在面对生成式 AI 训练这一新兴且复杂的场景时，上述路径均呈现出不同程度的不足，难以有效调和技术发展与著作权保护之间的矛盾。

3.1. 授权许可路径的市场性失效

授权许可路径，即要求使用者在使用作品前必须获得著作权人的事先许可。然而，这一路径在应对生成式 AI 训练数据的海量需求时，在市场中难以发挥作用，问题主要体现在以下几方面：

生成式 AI 数据训练对作品的使用具有规模巨大、权利主体高度分散的特点，这使得逐一获取事前授权在实践中几乎不具备可操作性。生成式 AI 数据训练的过程中，其功效发挥是基于拥有庞大参数量和规模化的数据库，而不是依靠少数作品，其数据体量远超传统作品使用场景[15]。在这种情况下，如果要求

开发者取得授权后才能进行数据训练，会产生的后果在于，大量的数据来源于网络爬虫等技术手段入侵信息系统[16]，因而开发者很难逐一定位到海量作品的著作权人的信息，就算能找到著作权人，使用作品数量巨大也对应着授权费用畸高。这种过高的搜寻成本与缔约成本，使得市场无法通过自愿交易实现资源配置。

即便训练者有意寻求授权，也面临着具体权利难以厘清的障碍。在生成式 AI 训练数据的过程中，其抓取的作品可能涉及著作权人的复制权、翻译权、汇编权等多项权利。然而真实的训练过程究竟涉及到哪些具体的权利，尚未有相对统一的认识，这就导致开发者难以向著作权人明确希望获得哪些权利的授权[17]。这种情况又进一步加剧了授权许可路径的实现困境，使开发者陷入欲授权而不能的境地。

由于授权许可在实践中困难重重，只有很少一部分作品能够获得授权，这又会导致模型偏见的问题。从数量上看，既然授权许可在生成式 AI 数据训练中难以执行，开发者难以满足数据合法收集的要求，那么可供使用的训练数据必然会大幅度的减少，这种不全面的样本可能会导致模型的偏见。从时间上看，授权许可的逻辑是先授权后使用，那么从授权到使用的过程可能时间较长，导致获得许可后的作品在内容上已经滞后于当下情况，不具备时效性，这也可能导致模型的偏见[18]。

3.2. 合理使用路径的适配性质疑

为了规避授权许可的高昂成本，部分学者试图将生成式 AI 训练行为纳入合理使用制度的豁免范围。然而，无论是基于我国《著作权法》的立法模式，还是国际通行的检验标准，合理使用路径在适配生成式 AI 训练场景时均面临着问题，主要体现在以下几方面：

从我国的法律规范来看，生成式 AI 数据训练难以落入我国《著作权法》第 24 条所列举的任一具体情形。该条款采取的“一般条件 + 具体情形”的立法模式，要求构成合理使用必须符合法定的 13 种情形之一。然而，在这 13 种情形中，和 AI 数据训练有关系的只有第一款的“个人学习”和第六款的“科学研究”，但数据训练行为与这两种情形并不适配。一方面，它不属于“个人学习、研究或者欣赏”，因为使用主体通常是商事主体而非自然人，且使用目的具有很强的营利性。另一方面，它也难以被归入“学校课堂教学或者科学研究”，因为 AI 对于作品的需求是海量、规模性的，已远超过这一款规定的“少量复制”的要求。此外，该条兜底条款的适用须以法律、行政法规的明确规定为前提，而目前并无此类规定。

即使尝试对我国的合理使用制度进行扩张解释，生成式 AI 训练也难以满足《伯尔尼公约》三步检验法这一判定标准。“三步检验法”是指在确定合理使用的范围时，应当考虑以下三个条件：1) 使用的目的是特殊情况；2) 使用的数量和范围是有限的；3) 使用不会对被使用作品的正常利用造成不合理的影响[18]。然而，数据训练行为并不能契合这一标准。一是数据训练作为一种广泛的商业性开发活动，并不具备特殊性；二是数据训练所使用的作品是海量的，范围也极为广泛，数量和范围并非有限；三是 AI 生成内容会对数据训练中使用的作品产生市场替代，对作品的正常使用显然造成了不合理的影响。

源于美国判例法的“转换性使用”理论，也不能很好的解释 AI 数据训练行为。“转换性使用”是指对他人作品进行使用时或具有再生产功能，或与著作权人对该作品内容的使用方式、目的或功能截然不同的作品使用行为，学者提出应把它作为合理使用判断四要素中“使用行为的目的和性质”考察的核心子要素之一，美国联邦最高法院在 Campbell 案中正式采纳对转换性使用因素的考察，主张“使用的转换性程度越大，合理使用分析中其他因素的重要性就越低”[19]。支持合理使用说的学者常借助转换性使用来论证训练行为的合理性，认为模型学习的是作品中的事实和思想而非独创性表达[20]。然而，这种观点简化了训练过程，AI 训练中的“转换”多为技术性的数据抽象处理，缺乏表达层面的创作性重构[15]。

3.3. 法定许可路径的制度性缺陷

法定许可制度作为一种折中方案，允许使用者在未经事先许可的情况下使用作品，但须依法支付报酬。这一制度看似能够兼顾效率与公平，但在应对生成式 AI 训练数据问题时，其固有的制度设计仍存在多处缺陷，难以直接适用。问题主要体现在以下几方面：

法定许可制度的价值取向与生成式 AI 数据训练的商业属性不匹配。一方面，我国《著作权法》所规定的法定许可情形，主要集中于编写出版教科书、报刊转载、广播组织播放录音制品等特定领域，其目的是支持教育、新闻传播、文化发展等非营利的公益性事业，而生成式 AI 即使可以促进技术的发展，也有一定的公共效益，但其具有明显的营利性特征，与法定许可的价值取向不相匹配。另一方面，法定许可要求使用的作品是“片段”、“短小”、“单幅”，而生成式 AI 对数据的大量抓取，已经超出了法定许可的预设框架。

法定许可的僵化的报酬机制无法适应 AI 数据训练的场景。我国法定许可的报酬确定标准主要有以版权计酬、以定额标准计酬和以使用比例计酬几种。如录制发行录音制品采用版税的方式付酬，即录音制品批发价 × 版税率 × 录音制品发行数。广播电台、电视台播放录音制品，可以与管理相关权利的著作权集体管理组织约定每年向著作权人支付固定数额的报酬。然而，这些标准都相对僵化，没有动态化的计酬标准，无论采取何种方式计酬，生成式 AI 的数据训练会持续不断地需要海量数据，其所面临的成本是一个巨大的挑战。

传统的法定许可制度未能充分尊重著作权人的意思自治，缺乏灵活的退出机制。在法定许可中，一旦作品发表且权利人未声明排除，使用者即可依法使用，权利人仅保留获酬权。这种安排客观上削弱了著作权人对作品传播的自主控制。很多创作者可能会因为担心市场份额被挤占而反对作品被用于 AI 训练，若完全剥夺其拒绝的权利，将严重损害其创作的积极性。

4. 准法定许可制度适用于生成式 AI 训练数据的理论证成

在厘清生成式 AI 训练数据面临的著作权困境与现行制度路径的局限性后，探索一种能够平衡技术创新与权利保护的例外路径便成为当务之急。准法定许可制度，要求生成式 AI 在使用已合法公开且未声明禁止使用的作品前进行公告，并设定异议期；著作权人可在异议期内提出反对以阻止使用，若无异议则服务提供者可依法使用并支付报酬。这一路径既能保障著作权人的获酬权和退出权，又可有效降低训练者使用作品的成本。

4.1. 准法定许可制度的法理正当性

准法定许可制度的法理正当性，源于其对知识产权制度本质的准确把握，以及与国家人工智能发展战略的高度适配。

准法定许可契合知识产权制度的本质。知识产权的根本目的并非是绝对保护私权利，而是通过赋予创作者有限专有权，保障其从智力成果中获得合理回报，进而激励持续创新，推动文化事业进步。准法定许可制度与这一核心目标相契合。它要求开发者向训练数据中作品的著作权人支付报酬，确保著作权人的经济利益。另外，它基于作品已公开发表的事实，推定著作权人对作品用于技术发展的一定容忍度，在充分尊重著作权的前提下，规避传统一对一授权模式，降低开发者的协商与交易成本，避免作品垄断，扩大作品传播范围。因而，准法定许可不仅不影响著作权激励功能的发挥，而且能够在人工智能产业与著作权人之间实现利益的平衡。

准法定许可符合国家人工智能发展战略。自 2017 年《新一代人工智能发展规划》提出培育具有国际竞争力的人工智能产业集群，到 2021 年《“十四五”数字经济发展规划》强调促进数据要素高效流通，

再到 2023 年《生成式人工智能服务管理暂行办法》明确合法获取数据、尊重知识产权的原则，我国已构建起清晰的人工智能产业的发展蓝图。准法定许可制度符合这一发展战略。它通过简化授权程序、降低成本，能够缓解开发者在数据获取上的压力，并且有利于数据要素的大规模流通与利用，呼应了促进数据要素高效流通的战略部署。因而，它既满足了产业发展的现实需求，符合国家发展战略，为生成式 AI 的可持续发展提供了坚实的法治保障。

4.2. 准法定许可制度的实践可行性

制度必须根植于现实，准法定许可制度并非空中楼阁，其在立法先例、技术支撑等方面均已具备坚实的实践基础，在实践中具有可行性。

我国相关立法已经做出了准法定许可制度的探索。《信息网络传播权保护条例》第九条已经构建了用于网络扶贫服务的准法定许可制度，做出的尝试主要包括：首先，摒弃了合理使用、法定许可等制度对著作权人的意思自治的排除，著作权人在公告期内享有禁止权，公告期后仍享有解除权。其次，克服了付酬标准的僵化，付酬标准并非法定，而是由网络服务者制定，且著作权人对此享有异议权。最后，省去复杂的授权环节，采取默示同意的规则，著作权人未明确表示拒绝即视为同意，提高了授权的效率[21]。生成式 AI 训练数据的准法定许可制度，在默示同意机制、公告异议的程序设计以及保障著作权人获酬权等方面，与上述制度存在高度相似，为其构建提供了可直接借鉴的立法先例。

现有技术体系可以支持准法定许可制度的运行。在公告环节，可依托区块链技术对作品使用信息进行不可篡改的记录与实时公示，以透明机制保障著作权人的知情权。在异议环节，通过数字水印与内容指纹技术可以精准溯源作品权属，并利用专有算法识别生成式 AI 所使用的作品，为法定响应期内的异议审查提供强大的技术验证支撑，从而提升争议解决效率与准确性。在付酬环节，智能合约系统可以基于标准化计费规则自动执行微支付结算，再配合区块链的可信账本机制，能够实现海量使用场景下的实时、精准收益分配，破解传统著作权交易中报酬支付迟滞与计算误差的困境[22]。

4.3. 准法定许可制度的优越性

相较于授权许可、合理使用与法定许可等既有路径，准法定许可制度展现出独特的优越性，集中体现在其能更好地兼顾效率与公平、协调产业发展与权益保护。

准法定许可制度能够更好地兼顾效率与公平。准法定许可制度在生成式 AI 的开发者与著作权人的利益之间划定了一个折中点。相较于授权许可模式，它允许开发者按照著作权人未明确表示拒绝即视为同意的方式使用作品，无需事前取得授权，降低了海量数据授权带来的时间成本，满足了技术发展的效率要求。相较于合理使用制度，它强制开发者付酬，确保了著作权人能够保留对其作品的最终控制权并获得经济补偿，兼顾了公平。相较于法定许可制度，它采取了灵活的计酬方式，且赋予著作权人异议权和退出权。因而相较于既有的三种路径，准法定许可制度能够更好地兼顾效率与公平。

准法定许可制度能够有效平衡产业发展与权益保护。从产业发展的角度来看，准法定许可能够保障生成式 AI 开发者以合理的成本合规使用作品，避免了因高昂成本而停止模型开发，规避了头部 AI 企业的数据垄断，能够促进市场的公平竞争与技术创新的多样性。从权益保护的角度看，对于著作权人而言，准法定许可制度通过前置的公告知情权、动态的异议退出权以及稳定的报酬请求权，为其构建了一套更具实效的利益机制。因此，准法定许可制度既能为开发者破解合规困境、释放创新活力，又能为著作权人提供制度化、可预期的权益保障渠道，具有显著的优越性。

5. 生成式 AI 训练数据准法定许可制度的系统构建

在证成准法定许可制度的正当性、可行性与优越性之后，如何将其从理论构想转化为具体可操作的

法律规则，便成为制度落地的关键。将准法定许可制度适用于生成式 AI 数据训练，必须通过精密的权利义务配置、清晰的程序设计与健全的配套机制，为生成式 AI 产业的健康发展与著作权人的权益保障提供坚实的法治基础。

5.1. 主体的权利与义务配置

在生成式 AI 数据训练的准法定许可制度中，涉及到两方主体，对数据训练使用作品享有著作权的著作权人是权利主体，而生成式 AI 的开发者为义务主体。适用准法定许可制度，需要明确他们的权利义务。

对于著作权人而言，其权利体系应当实现从事前防范到事后救济的全覆盖，以改变其在数据使用中的弱势地位。首先，应赋予著作权人保留声明权，即允许著作权人在作品发表时或发表后，以明确的方式声明禁止其作品被用于 AI 训练。其次，在作品被爬虫抓取后，著作权人享有知情权与异议权。生成式 AI 开发者必须进行公告，使著作权人能够知悉其作品被使用的具体情况，在公告期内，著作权人若不同意使用，可行使异议权以阻断该使用行为，如果对公告的付酬标准不满意，亦可行使异议权。最后，即便作品已过公告期并被使用，著作权人仍应享有报酬请求权与动态的删除请求权。作品被使用后，著作权人有权要求生成式 AI 开发者按照公告的标准支付报酬，并且可以随时要求开发者停止使用作品，同时要求对其作品彻底删除。即使作品在使用后被删除，也可以要求开发者按照公告标准支付使用期间的报酬。

对于生成式 AI 开发者而言，在获得授权的同时，必须承担起一系列严格的义务。首先，开发者最基础的义务是信息披露义务，生成式 AI 抓取数据前，应当以标准化的格式，全面、准确地公告训练数据中所含作品清单、著作权人信息及报酬标准，以保障著作权人的知情权与异议权。其次，开发者应加强算法的透明度，为了避免著作权人对自己作品产生失控感，开发者应公开算法的功能特性、工作原理、运行逻辑以及隐藏的风险，以增进著作权人对 AI 的了解以及自己作品的去向，从而增强对生成式 AI 技术的信任，增强其愿意将作品用于 AI 数据训练的意愿。最后，开发者应当保障在作品使用的过程能够响应著作权人删除作品的请求。开发者可以建立用户数据标记系统，对作品附加可追溯标记，当著作权人要求删除时，系统即自动停止后续的使用并删除已存储的原始数据。

5.2. 程序的构建与运行保障

权利的实现有赖于公正、高效的程序设计。准法定许可制度的核心程序在于构建一个以公告与异议为主线的实施框架。

生成式 AI 在使用作品进行数据训练前，必须进行明确的公告。公告内容应法定化，至少涵盖生成式 AI 开发者的身份、拟使用的作品信息、明晰的报酬计算方案以及异议提交的渠道。公告形式采取可供检索的网络公告模式，对于不同类型的作品分类进行公告，依托知识产权公共服务平台进行集中统一发布，以确保信息的公信力与可检索性。公告期限需长短适宜，可借鉴《信息网络传播权保护条例》第 9 条设定的 30 日期限。

在异议处理环节，应建立高效的响应流程。为保障著作权人异议权的行使，应当由负责公告的知识产权公共服务平台负责处理异议，著作权人可通过该平台提出异议，开发者也应在该平台上同步处理进度和结果，保证与著作权人的信息同步与交流。对于公告期内的异议，开发者应当立即将相关作品从训练数据中移除，并通过标记等方式确保不再抓取。对于公告期之后的异议，开发者不仅要进行删除，还应按照公告的付费标准对已经使用的作品支付使用期间的报酬。

在具体实施过程中，著作权行政主管部门应当建立常态化的监督检查机制，定期对开发者的公告合规性、异议响应及时性以及报酬支付准确性进行评估，并对违规行为实施相应的行政处罚。具体而言，

需要建立程序运行的评估与优化机制，定期对公告平台的访问量、异议处理效率、报酬支付准确率等指标进行监测评估，及时发现和解决程序运行中存在的问题。同时，建立用户反馈机制，广泛听取著作权人和开发者对程序运行的意见建议，持续优化程序设计，提升用户体验。通过这种动态调整的机制，确保准法定许可程序能够适应技术发展和实践需求的变化，始终保持其有效性和适用性。

5.3. 制度的配套与衔接机制

准法定许可制度的顺畅运行，需要科学的配套机制予以支撑，并与现有法律体系实现有机衔接。

构建一个灵活、科学的报酬定价体系是维系准法定许可制度运行的关键。可以在著作权行政管理部门的指导下，由行业协会、知识产权保护中心、著作权人代表等主体，综合考虑作品类型、市场热度、独创性高度等因素，通过民主协商的方式确定作品的付酬标准。在这个过程中，行政管理部门要做好引导，使著作权人和开发者把握准法定许可的小额、普惠、便捷的特点，合理确定报酬，防止标准过高而导致开发者承担过重的成本。此外，随着技术的发展，原有的计酬标准可能与现实情形不符，行政管理部门应随时监测标准是否可靠，出现这种情况时召集多方主体再次进行协商。

监管体系与纠纷解决机制的多元化构建，是维护制度权威与公信力的保障。在监管层面，应形成以知识产权行政管理部门为主导，网信、工信等部门协同配合的综合治理格局。在纠纷解决层面，鉴于训练数据纠纷具有涉众面广、单案标的额可能较小的特点，必须突破单一诉讼路径，发展多元化解机制，应鼓励和支持著作权集体管理组织发挥其专业优势，代为收转报酬，并先行调解相关争议。另一方面，可设立快捷的行政裁决通道，并积极引入仲裁、行业调解等替代性纠纷解决机制，形成诉讼、行政处理、社会化解纷方式优势互补的多元共治体系。

6. 结语

生成式 AI 作为数字时代的创新引擎，其发展与著作权保护的平衡是科技进步与法治建设必须回应的时代命题。海量高质量训练数据的需求与现有著作权制度的冲突，既制约着 AI 产业的创新活力，也考验着知识产权保护的实际效果。授权许可的实操困境、合理使用的适配局限与法定许可的制度缺陷，证明传统路径已难以应对新技术带来的挑战，亟需构建兼顾效率与公平的新型规制方案。准法定许可制度立足知识产权制度本质与国家 AI 发展战略，既破解了开发者的授权难题与成本压力，又赋予著作权人事前防范、事中异议、事后获酬的全链条权利保障，实现了产业发展与权益保护的动态平衡。该制度依托既有立法先例与成熟技术支撑，具备充分的法理正当性与实践可行性，为化解 AI 训练数据的著作权困境提供了切实可行的路径。

准法定许可制度的落地，需通过清晰的权利义务配置、规范的程序设计与多元的配套机制，将理论构想转化为可操作的法律规则。未来，还需随着技术迭代与实践发展，持续优化报酬定价体系、完善监管机制与纠纷解决路径，推动制度与现有法律体系深度衔接。唯有如此，才能为生成式 AI 产业的健康发展筑牢法治根基，在激励创新与保护权利之间实现有机统一，助力我国在全球人工智能竞争中抢占先机，推动新质生产力与文化繁荣的协同发展。

参考文献

[1] 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): 80-88.

[2] 吴汉东. 人工智能生成作品的著作权法之问[J]. 中外法学, 2020, 32(3): 653-673.

[3] 游涛, 计莉卉. 使用网络爬虫获取数据行为的刑事责任认定——以“晟品公司”非法获取计算机信息系统数据罪为视角[J]. 法律适用, 2019(10): 3-10.

[4] 张吉豫, 汪赛飞. 大模型数据训练中的著作权合理使用研究[J]. 华东政法大学学报, 2024, 27(4): 20-33.

[5] 王健, 吴宗泽. 生成式人工智能反垄断论纲[J]. 法治研究, 2024(6): 130-147.
[6] 孙济民. 从合理使用到合理学习: 人工智能训练数据的著作权困境与规范重构[J]. 中国法律评论, 2025(5): 52-65.
[7] 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.
[8] 韩雨潇. 人工智能大模型训练数据的版权风险与化解路径[J]. 中国出版, 2025(2): 54-59.
[9] 黄孟苏. AI 大模型训练数据的版权风险与治理路径[J]. 湖北大学学报(哲学社会科学版), 2025, 52(5): 185-193.
[10] 张平. 人工智能生成内容著作权合法性的制度难题及其解决路径[J]. 法律科学(西北政法大学学报), 2024, 42(3): 18-31.
[11] 戴文怡, 肖冬梅. 生成式人工智能训练数据的著作权法因应: 著作权合规方案[J]. 图书情报知识, 2025, 42(1): 89-100.
[12] 魏丽丽. 生成式人工智能数据训练中著作权限制制度的适用选择[J]. 政法论丛, 2025(5): 159-172.
[13] 刘一帆. 生成式人工智能数据训练行为的著作权困境与出路[J]. 江海学刊, 2025(4): 182-189+256.
[14] 任保平, 郭晗. 新科技革命背景下形成新质生产力的战略逻辑与实践路径[J]. 商业经济与管理, 2024(8): 21-29.
[15] 王文敏. 人工智能对著作权限制与例外规则的挑战与应对[J]. 法律适用, 2022(11): 152-162.
[16] 陈锐, 江奕辉. 生成式 AI 的治理研究: 以 ChatGPT 为例[J]. 科学学研究, 2024, 42(1): 21-30.
[17] 陈雨悦, 李军. 生成式人工智能数据训练的著作权困境及其对策[J]. 时代法学, 2025, 23(1): 58-79.
[18] 张涛. 生成式人工智能训练数据集的法律风险与包容审慎规制[J]. 比较法研究, 2024(4): 86-103.
[19] 李杨. 著作权侵权认定中的转换性使用理论适用阐释[J]. 北方法学, 2023, 17(3): 42-56.
[20] 姚志伟, 黄丹敏. 人工智能训练中作品的合理使用问题[J]. 西安交通大学学报(社会科学版), 2025, 45(6): 121-130.
[21] 侯玉玲. 图书馆开展网络扶贫服务思考——以《信息网络传播权保护条例》第九条为依据[J]. 图书馆研究, 2015, 45(6): 82-85.
[22] 赵立冬. 合理使用还是法定许可? 生成式人工智能训练数据著作权规制例外路径研究[J/OL]. 图书馆建设, 1-15. https://link.cnki.net/urlid/23.1331.G2.20250917.1316.002, 2026-03-28.