

生成式人工智能语料训练的知识产权 监管制度：域外经验及其启示

蒋洁如 林 鹤*

(上海图书馆(上海科学技术情报研究所), 上海 200031)

摘要: [目的/意义] 生成式人工智能技术在语料训练阶段会触发著作权侵权风险, 目前全球范围已经出现了侵权诉讼案件。提出应对著作权侵权风险的中国监管方案, 对平衡我国人工智能与传统内容创作产业的利益关系, 推动人工智能技术可持续发展有着重要意义。[方法/过程] 通过文献、比较和经验分析, 揭示了发达国家(地区)的监管制度。在借鉴域外监管经验的基础上, 结合我国国情提出监管方案。[结果/结论] 研究发现, 我国生成式人工智能技术处在发展阶段, 政府应推出灵活的行政监管原则, 在著作权法律体系中构建适用于生成式人工智能的柔性的“合理使用”制度, 同时制定适用于生成式人工智能的知识共享协议方案, 作为节约许可成本的行业合规路径。

关键词: 生成式人工智能; 语料训练; 知识产权; 著作权; 监管制度; 美国; 欧盟; 英国; 日本; 中国; 合理使用; CC 协议

DOI: 10.3969/j.issn.1008-0821.2025.06.009

[中图分类号] TP18; D923.41 [文献标识码] A [文章编号] 1008-0821 (2025) 06-0102-10

The Intellectual Property Regulatory Systems of Generative Artificial Intelligence Data Training: Extraterritorial Experience and Enlightenment

Jiang Jieru Lin He*

(Shanghai Library(Institute of Scientific and Technical Information of Shanghai), Shanghai 200031, China)

Abstract: [Purpose/Significance] The generative artificial intelligence technology has triggered the copyright infringement risk during the generative AI data training stage, and infringement litigation cases around the world have emerged. It is of great significance to propose regulatory and compliance measures to deal with the copyright infringement risk of generative AI data training in China, to balance the interests of AI industry and traditional content creation industry, and promote the sustainable development of GenAI technology. [Method/Process] Through the method of literature, comparison and empirical analysis, this paper revealed the regulatory systems of developed countries(areas). On the basis of comparing and learning from the regulatory experience of developed countries, this paper gave the countermeasures in combination with our national conditions. [Result/Conclusion] The study found that China is at the rapid stage of generative AI development. Chinese government should introduce adaptable administrative regulatory principles. Meanwhile, China should incorporate a flexible “fair use” system that is applicable for GenAI data training into the current copyright legal system. At the same time, the industry can also develop a Creative Commons License scheme applicable to GenAI data training as a cost saving compliance measure.

Key words: generative artificial intelligence; data training; intellectual property; copyright; regulatory system; U. S. A.; European Union; U. K.; Japan; China; fair use; CC license

收稿日期: 2024-12-03

作者简介: 蒋洁如 (1990-), 女, 助理研究员, 研究方向: 知识产权法、人工智能。

通信作者: 林鹤 (1975-), 男, 咨询研究中心副主任, 高级工程师, 研究方向: 情报分析、知识库和行业信息化、知识产权。

OpenAI ChatGPT 的诞生掀起了生成式人工智能 (Generative Artificial Intelligence, 以下简称“GenAI”) 的发展浪潮。进入 2024 年, 海外大模型持续迭代, OpenAI ChatGPT 已经更新到了 GPT-4o 版本, 与谷歌的 Gemini 和 Meta AI 的 LLaMA 大模型共同成为全球大模型发展的重要推动力。与此同时, 国产大模型走向成熟, 中国已经涌现出了 DeepSeek、文心一言、通义千问等 300 多个大模型产品^[1], GenAI 呈现百花齐放的状态, 加速赋能医疗、金融、制造等各个行业。随着 GenAI 技术的发展, GenAI 语料训练行为引发了知识产权风险, 并且已经在美国、英国等地出现了知识产权侵权诉讼案例。我国 GenAI 技术发展迅速, 出现大规模的知识产权侵权诉讼只是时间问题, 然而, 我国目前对于 GenAI 语料训练阶段数据使用的知识产权侵权问题仍然存在监管空白。本文通过比较美国、欧盟、英国和日本的知识产权监管制度, 结合我国的实际国情, 提出我国构建 GenAI 语料训练知识产权监管制度的可行方案, 以期最大程度地平衡 GenAI 产业和传统内容创作产业的利益关系。

1 GenAI 及语料训练的概念

GenAI 是一种能够创造新内容的人工智能系统, 通过深度学习模型学习输入的数据集 (语料库) 的模式和结构, 经过训练之后, 形成输出多模态内容的能力。一旦系统接收到人工输入的提示词, 就会触发一系列的复杂计算, 然后产生相较训练数据集具有一定创新程度的多模态新内容, 包括文字、图片、音视频等^[2]。总之, GenAI 通过对大规模的语料库形成的训练数据进行学习, 不断地优化模型策略, 从而创造出基于原始数据的新的关联内容。

数据集是 GenAI 的重要组成部分, 数据集包括文本、图片、音视频等多模态数据, 这些数据可能来自网络中的书籍、文章、图片、音像、数据库、代码等公开资源, 也可能来自企业或个人的私有数据的贡献, 通过对多模态数据集进行数据加工形成可以用于训练模型的语料。GenAI 语料训练的整个流程包括: ①大规模多模态的数据收集: 包括从公开或非公开渠道收集大规模的多模态数据, 数据量常达数万亿级别。②数据预处理: 包括对收集的数据进行清洗、对齐、标注等处理, 并通过统一的格式进行标准化, 形成可用于训练的高质量语料数据

集。③数据预训练: 根据训练的规模和预训练任务的特点, 选择合适的大模型架构, 在预处理后的大规模语料数据集上进行无监督训练, 该过程需大量的计算资源和时间。④模型微调和优化: 为适应不同的任务需求, 还会基于少量标注数据进行有监督微调以适配具体任务, 借助人类反馈强化学习优化输出, 进而提升模型综合性能。GenAI 语料训练是一个持续投喂数据、学习和优化的动态过程。

2 GenAI 语料训练的知识产权风险

GenAI 语料训练已经引发了一系列的知识产权诉讼案例, 主要围绕著作权侵权争议展开。最典型的当数美国《纽约时报》与 OpenAI 之间的著作权侵权诉讼。美国《纽约时报》向 OpenAI 及其投资人微软公司提起指控, 指控二者未经许可使用其数百万篇文章来训练 ChatGPT 等 GenAI, 侵犯了《纽约时报》对这些文章报道享有的版权, 训练行为也使 GenAI 成为可靠的信息来源, 与新闻机构形成竞争关系。这项起诉拉开了未经许可使用已发布作品来训练人工智能技术的“法律战役”的新篇章^[3]。

版权法保护具有独创性的作品, 无论是英美法系国家, 还是大陆法系国家, 其版权法多数对独创性的要求一般只需要符合“作者独立创作”和“最低限度的创造性”两方面的条件即可^[4-5]。GenAI 的语料来源包括文本、图片、视频、音频等, 大部分数据不是专门为训练而创建的, 而是从互联网抓取的公开内容, 这些公开内容的绝大多数是符合独创性要求的“作品”。因此, GenAI 的语料数据集往往包含了拥有版权的作品。在模型训练之前, 需要对海量的语料数据集进行加工形成训练数据集, 包括数据下载、清洗、去噪、标引等加工步骤, 往往涉及对语料数据的复制、修改、汇总、删除等数据使用行为。根据《中华人民共和国著作权法》的规定, 著作权人享有对作品的复制权、改编权、翻译权和汇编权等专有权利, 未经著作权人许可对作品进行复制、改编、翻译或者汇编, 将构成著作权侵权。因此, 如果未经权利人的许可, 对受到著作权保护的语料内容进行复制、修改、汇总或删除, 这一系列的数据使用行为很可能会因为侵犯了著作权人的复制权、改编权、汇编权等专有权利而产生侵权风险^[6]。

2023 年以来, 全球已发生数十起作家、新闻媒

体等版权方起诉 OpenAI、Google 等 GenAI 公司未经许可使用其作品进行 GenAI 语料训练的版权侵权案件,我国也已出现首例小红书“TriK” GenAI 语料训练著作权侵权纠纷。这些诉讼呈现三大特点。第一,诉讼主体:主要发生在新闻媒体、出版商、作家等传统内容创作者与 GenAI 开发公司之间。第二,跨国性:由于一些 GenAI 公司为跨国公司,需应对多国(地区)提起的著作权侵权诉讼。第三,争议焦点:原告方主张未经许可的 GenAI 语料训练的数据使用行为构成版权侵权,具体侵犯了内容创作者对作品享有的复制权,而被告方则抗辩其行为属于版权法体系下的“合理使用”。

上述案件的核心争议点在于未经许可的 GenAI 语料训练是否构成版权侵权,以及是否适用版权法的“合理使用”制度。尽管这些案件仍在审理中,但美国、欧盟等发达国家(地区)的知识产权监管制度已为相关争议提供了一定解答依据。然而,我国对 GenAI 语料训练是否构成版权侵权、侵权边界以及是否适用“合理使用”等问题存在监管空白。因此,我国需尽快建立适用于 GenAI 语料训练的知识产权监管制度,并借鉴国际经验,提出既与国际接轨又符合国情的监管举措,以应对全球治理协同趋势下,跨国性知识产权纠纷带来的挑战。

3 域外 GenAI 语料训练的知识产权监管制度

3.1 美国:先发展后监管,依“四要素”判定 GenAI 是否落入“合理使用”

美国是全球 GenAI 技术的引领者,不仅拥有 OpenAI、Anthropic 等代表性的初创企业和微软、Google 等巨头,还拥有斯坦福人工智能实验室、麻省理工计算机科学与人工智能实验室等顶尖的研发机构。2023 年,美国在人工智能领域的私人投资总额达 672 亿美元,是中国的 8.7 倍,其中生成式人工智能投资总额达 252 亿美元,较 2022 年显著增长^[7]。

美国也是发生 GenAI 语料训练知识产权侵权诉讼最多的国家,但美国倾向于奉行“发展优先,监管后置”的原则。行政监管方面,美国提出框架性的监管指导原则来治理 GenAI。美国拜登政府曾在 2023 年签署了首个涉及生成式人工智能监管的行政法令(Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence)^[8],

提出八项监管指导原则,包括:确保安全可靠、确保公平公正实施、促进创新和竞争、加强美国的海外 AI 领导地位等。不过,针对性的 GenAI 行政监管方案尚未推出。

美国在 GenAI 版权法律制度方面采取审慎态度,国会明确表示暂不进行专门立法,而是依托现有法律框架处理相关问题^[9]。作为判例法国家,美国法院现阶段主要依据“四要素”分析法来判定人工智能语料训练或类似的计算机文本与数据挖掘行为是否构成“合理使用”的侵权例外。美国《版权法》第 107 条确立了“合理使用”的四要素分析法,分别是:①使用的意图和特征,包括使用是否具有商业性质或用于非营利性教育目的。②版权作品的自然属性。③使用版权作品的数量和实质使用的比例。④使用行为对版权作品价值的潜在市场影响^[10]。司法实践中,第一要素(使用目的及性质)和第四要素(潜在市场影响)的判定占据更大的权重^[11]。在 Thomson Reuters et al v. ROSS Intelligence 一案中,法院判定罗斯公司(ROSS Intelligence)的 AI 训练行为不构成“合理使用”。法院在第一要素判定中,认为罗斯公司使用汤森路透旗下的 Westlaw 数据库的案件摘要来训练其 AI 法律研究搜索引擎模型(非 GenAI)的数据使用不具有“高度转换性”,因为其最终目的是创建与 Westlaw 直接竞争的法律研究工具,这与汤森路透使用案件摘要的目的极度相似。法院在第四要素判定中,认为罗斯 AI 语料训练的数据使用行为对 Westlaw 的市场产生了负面影响,罗斯开发的 AI 工具与 Westlaw 构成竞争,且有影响汤森路透的潜市场(如数据市场)的可能^[12]。相反,在 Authors Guild v. Google 案中,法院认定谷歌的计算机文本与数据挖掘行为构成合理使用,因为其具有高度转换性,即通过建立搜索功能为作品提供新的使用方式,未对原作品市场形成替代,反而促进了知识的传播^[13]。

3.2 欧盟:实施强监管,版权法“合理使用”条件严苛

欧盟充分认识到 GenAI 语料训练引发的知识产权风险,与美国的“观望”态度不同,欧盟国家实行了“强监管”。欧盟议会通过了世界上首个全面且严格的人工智能监管法规——《欧盟人工智能法案》(EU AI Act)^[14],法案将 GenAI 语料训练的知

识产权问题与欧盟现有的版权制度进行了衔接，法案明确规定，如果语料训练涉及版权受保护内容的数据使用，需要经过版权权利人的授权许可，除非适用《数字化单一市场版权指令》（以下简称《版权指令》）的“合理使用”规定。

《版权指令》要求计算机文本与数据挖掘行为满足“合理使用”的条件十分严苛^[15]。《版权指令》第 3 条为文本与数据挖掘的“合理使用”设定了限制条件：在行为主体方面，《版权指令》将其定义为“科研机构和文化遗产机构”，但是否可以包含商业性研究机构则由欧盟成员国国内法自行决定；在行为目的方面，仅限定为科学研究。也就是说，只有科研文化机构以科学研究为目的的 GenAI 训练行为才能够落入“合理使用”的范畴。此外，《版权指令》第 4 条指出，如果权利人明确保留了他们的权利以阻止被计算机文本和数据挖掘，则不适用于“合理使用”的规定。

值得注意的是，欧洲近年来在人工智能领域的发展速度落后于美国和中国，不论是私人投资还是政府投资都少于美国和中国。欧盟强监管的实施也遭到了社会各界的批评，认为欧盟在生成式人工智能的起步阶段对其实施过度监管是错误的，这会阻碍人工智能企业在欧洲的发展，可能导致欧洲在这个领域持续落后于美国和中国^[16]。

3.3 英国：主张灵活监管，拟扩展版权法“合理使用”至商业用途

英国正在积极争夺欧洲的人工智能领导地位。英国人工智能市场扩张几乎是法国、德国和欧洲其他国家总和的两倍，预计到 2035 年，人工智能市场将为英国经济增加 1 万亿美元^[17]。英国除了在人工智能领域拥有如 DeepMind、Albrain、StabilityAI 等著名的人工智能公司之外，还拥有一大批全球领先的高校院所，如伦敦大学学院、国王学院和帝国理工学院，使得英国能更好地在前沿研究和商业应用之间建立联系以推动产业发展。

英国本土已经出现了 GenAI 语料训练的版权侵权纠纷^[18]，但英国现行的版权法案的“合理使用”制度对人工智能企业不利。英国《版权、设计与专利法案》（CDPA）第 29A 条规定了计算机文本与数据挖掘行为的侵权例外，仅允许研究人员复制任何受版权保护的材料用于计算分析^[19]。也就是说，英

国版权法案的侵权例外规定只允许是以非商业性研究为目的的计算机文本和数据挖掘行为，且仅用于复制数据的行为。英国政府认为，对 GenAI 企业实施严格和繁重的法律规制会阻碍技术创新，并且意识到了现行版权法案仅支持非商业用途的“合理使用”的规定无法推动 GenAI 技术发展的的问题，因此，多次提出修法提案，试图将侵权例外扩展至商业目的^[20]。英国政府曾在 2022 年提议，只要 AI 开发者通过许可证、订阅或条款许可合法访问作品，就可以出于商业目的进行文本和数据挖掘，然而，这一提案遭到了内容创作产业的反对而被迫放弃^[21]。最近，其再一次提出了将侵权例外扩展至处于商业目的的修法提案，并考虑引入透明度要求，比如要求 AI 开发者披露用于训练的数据来源^[22]。

行政监管方面，英国政府通过灵活的监管方式来解决人工智能行业与传统内容创作行业的利益矛盾。其在 2023 年发布的《创新的人工智能监管方式》白皮书（A Pro-innovation Approach to AI Regulation）^[23]中，提出基于五项原则对人工智能产业实施监管，分别是：安全、可靠、稳健；适当的透明度和可解释性；公平；问责制和治理；可竞争性和补救。各个监管部门将在五项原则的指导下，根据使用人工智能的具体环境来协同实施。此外，白皮书也指出，政府将确保该监管框架在保护版权的前提下，允许使用受版权保护的数据训练 AI 模型。

3.4 日本：对“合理使用”制度提出适用于 GenAI 的宽松解释

日本相比其他经济体，没有特别出色的人工智能企业，在这一轮的 GenAI 竞争中跌出了第一梯队，但日本正在奋起直追，日本政府计划斥资约 65 亿美元来发展半导体芯片和 AI 产业^[24]。2024 年，日本 GenAI 市场规模约为 45 亿美元，同比增长率为 35.5%，预计到 2027 年将增长到约 73 亿美元^[25]。

在日本大力发展 GenAI 的背景下，日本政府曾发布《AI 业务指南》，提出通过灵活的监管框架来支持人工智能技术的发展，初期以鼓励创新为主，随着技术成熟逐步引入更严格的监管要求。该指南提出了包括以人为本、安全、公平、隐私保护等十项监管原则，并且进一步提出各方要遵循确保安全、隐私保护的原则来确保妥善处理 AI 训练的知识产权问题^[26]。

法律制度方面，日本新修的《著作权法》允许将受著作权保护的作品用于计算机对数据的提取、比较、分类或其他统计分析，除非是以“享受为目的”的数据分析行为^[27-28]。2024 年 3 月，日本文化厅版权办公室专门发布了《关于人工智能和著作权的一般理解》的文件(General Understanding on AI and Copyright in Japan)^[29]，对 GenAI 语料训练的数据使用行为如何适用于现行《著作权法》第 30-4 条的“合理使用”条款进行了明确解释。GenAI 语料训练如果是“非享受作品所表达的思想或情感

为目的”的使用，即便未经权利人的许可，也是被允许的。这里的“享受”是指通过使用受著作权保护的作品，使观众的智力和情感需求得到满足，从而获得利益的行为。著作权权利人通过作品获得的经济利益通常被认为是因满足观众智力和情感需求而获得奖励，而以非享受目的使用作品，一般被认为不会损害著作权权利人的经济利益^[30]。

3.5 域外制度对比总结

上述地域有关 GenAI 语料训练的知识产权监管制度对比如表 1 所示。

表 1 各国(地区)GenAI 语料训练知识产权监管制度对比
Tab. 1 Comparison of GenAI Data Training IP Regulations in Different Countries(Regions)

技术实力与 监管制度	国 家				
	美 国	欧 盟	英 国	日 本	中 国
技术实力	最强	较弱，发展趋缓	较强，发展迅速	较弱，发展迅速	较强，发展迅速
行政监管	遵循八项监管指导原则，未推出针对性的行政监管方案	遵循欧盟人工智能法案：未经版权人许可不得使用受版权保护内容的数据进行语料训练	遵循五项监管指导原则，政府将采取行动确保监管体系在保护版权的前提下，允许使用受版权保护的数据训练 AI	遵循灵活的监管框架，各方要遵循确保安全、隐私保护的原则来确保妥善处理 AI 训练的知识产权问题	GenAI 语料训练不得侵害他人知识产权
法律监管	现行《版权法》依“四要素”判定 GenAI 是否落入“合理使用”	现行《版权指令》仅允许科研文化机构以科学研究为目的的 GenAI 训练行为落入“合理使用”	现行《版权、设计与专利法案》仅允许非商业目的的 GenAI 语料训练行为落入“合理使用”，政府拟通过修法扩大至商业目的	现行《著作权法》允许以“非享受为目的”的 GenAI 语料训练落入“合理使用”	现行《著作权法》的“合理使用”制度不适用于 GenAI 语料训练
监管基调	灵活，相对宽松	严苛	趋向宽松	灵活，相对宽松	不明确

在行政监管方面，美国、英国和日本均通过灵活的监管原则来指导应对 GenAI 的知识产权问题。英国提出以五项监管原则为核心的行政监管体系，要在保护版权的前提下，允许使用受版权保护的数据训练 AI 模型。日本提出在十项监管原则指导下，各方要遵循确保安全、隐私保护的原则来确保妥善处理 AI 训练的知识产权问题。欧盟则实施强监管，通过《人工智能法案》，将 GenAI 语料训练的知识产权问题与现行的《版权指令》进行了衔接，提出未经版权人许可不得使用受版权保护内容的数据进行语料训练。

在法律监管方面，上述地域的现行版权法的

“合理使用”制度对于 GenAI 的语料训练的数据使用行为均具有不同程度的适用性。美、日两国在 GenAI 语料训练的“合理使用”认定上采取了不同的弹性标准。美国采用“四要素”分析法，并进一步采用了“转换性使用”的判断规则，即通过考察使用行为是否赋予原作新的目的或表达，评估其是否对原作市场形成了替代。转换性越强，对原作市场的冲击越小，越可能构成合理使用。日本则以“非享受性使用”为标准，重点判断使用行为是否影响了著作权人通过作品获取满足观众的智力情感的收益的权利，若语料训练不涉及作品的“享受性”使用，则可能被认定为合理使用。这两种制度设计均体现了

在保护著作权的同时，为技术创新预留必要空间的立法智慧。相比之下，欧盟和英国采取了相对严格的立法。欧盟现行版权法将 GenAI 训练的合理使用范围严格限定于科研文化机构的非商业性研究目的。英国虽同样采取仅允许非商业目的的 GenAI 训练行为的立场，但英国政府已启动修法程序，拟将合理使用制度扩展至商业性行为，其立法动态显示出适度放宽的趋势。

总体来说，美国和日本关于 GenAI 的语料训练的整体知识产权监管环境相对宽松，英国则趋向宽松，而欧盟的监管环境最为严苛。在不同的监管基调之下，美国的技术竞争力最强，英国和日本的 GenAI 技术发展迅速，而欧盟整体的 GenAI 技术发展趋缓，且落后于中、美。

4 域外知识产权监管制度对我国的启示

4.1 设立灵活的行政监管框架，推动 GenAI 技术创新发展

美日两国采取了相对宽松的政策导向，英国也呈现出宽松趋势，而欧盟则实施了最为严格的监管框架。这种差异化的监管环境影响了各国（地区）的技术发展态势。凭借其相对灵活和宽松的政策环境，美国保持了显著的技术竞争优势，英国、日本也实现了 GenAI 技术的飞速发展；相比之下，欧盟严格的监管制度在一定程度上制约了其技术进步，导致发展速度滞后。可见，在技术发展初期实施强监管很可能会阻碍人工智能技术的发展。从各国（地区）行政监管的角度来看，美国、英国和日本均通过灵活的行政监管框架来构建相对宽松的监管环境，英国和日本提出了安全、公平、隐私保护等监管指导原则，在这些监管原则的指导下，各监管方协调合作来确保妥善处理 AI 训练的知识产权问题。

国家互联网信息办公室公布了《生成式人工智能服务管理暂行办法》^[31]，提出 GenAI 预训练、优化训练等训练数据处理活动不得侵害他人知识产权。全国网络安全标准化技术委员会发布了《网络安全技术 生成式人工智能服务安全基本要求》标准文件^[32]，要求“语料用于训练前，应对语料中的主要知识产权侵权风险进行识别，发现存在知识产权侵权等问题的，服务提供者不应使用相关语料进行训练；语料中包含作品的，应重点识别语料以及生成

内容中的著作权侵权问题。”上述行政监管办法已经认识到语料训练的知识产权侵权问题，然而，对于使用作品数据的语料训练行为是否一定会构成著作权侵权的问题，均未给出明确解释，并且，使用“不得侵犯”这样绝对化的规范表述，呈现出趋向严苛的监管基调。

我国正处于 GenAI 技术的高速发展阶段，在全球人工智能创新领域彰显出显著竞争力。基于当前技术发展态势，建立灵活适度的监管框架更符合产业发展需求。建议相关行政部门在制定监管政策时，可参考英、日等国的实践经验，以“安全可控、公平有序、透明可信”为核心原则，构建协同监管机制，审慎处理 GenAI 语料训练中的知识产权问题。鉴于 GenAI 技术具有迭代迅速、应用场景多元、跨领域融合等特征，采取指导性的监管原则能够实现动态适应性监管，有效协调多部门监管职能。同时，建议建立多方参与的监管评估机制，集合学术界、AI 企业、内容创作者等利益相关方，对监管政策实施效果进行持续监测与评估，及时优化调整监管框架。这种协同治理模式不仅能够保持监管政策的科学性与合理性，更能有效平衡 GenAI 创新企业与传统内容创作产业之间的利益关系，推动产业生态的良性发展。

4.2 完善适用 GenAI 的著作权法律制度，设定“合理使用”柔性边界

“合理使用”是著作权法律体系中的一项重要侵权例外制度。根据“合理使用”制度，在特定条件下，不经过著作权人的许可且不向其支付报酬，使用著作权人已经发表的作品，不视为侵权。美、欧、英、日现行版权法的“合理使用”制度对于 GenAI 语料训练行为均具有适用性。然而，我国现行《中华人民共和国著作权法》（以下简称《著作权法》）的“合理使用”制度不适用 GenAI 语料训练。我国《著作权法》明确列举了落入“合理使用”的 12 种行为^[33]，包括以个人学习、研究或者欣赏为目的的使用；以教学或者科学研究为目的的少量复制行为等。GenAI 语料训练的数据使用行为涉及对数据的大量复制，且绝大多数以商业盈利为目的，因此，无法被《著作权法》的 12 种合理使用行为所囊括。

从《著作权法》的制度理性出发,将未经许可的作品数据训练行为一概视为侵权显然缺乏合理性。这种绝对化的认定标准将导致 GenAI 开发者面临难以承受的许可成本。在数万亿级别的训练数据集中,开发者需要识别并获取海量权利人的授权许可,这不仅会严重阻碍技术创新,更可能导致 GenAI 市场的系统性失灵。基于此,建议在我国现行《著作权法》框架下构建适用于 GenAI 训练数据的“合理使用”制度。这一制度设计具有双重价值:一方面,在保障著作权人合法权益的同时,为 GenAI 技术创新提供必要的制度空间,促进产业健康发展;另一方面,与国际主流立法趋势相协调,实现与美、欧、英、日等发达国家和地区法律制度的有效衔接。这种平衡保护机制既符合数字时代著作权法的立法宗旨,也顺应了全球人工智能产业的发展需求。

4.2.1 构建 GenAI “合理使用”制度的合法基础

我国现行《著作权法》对“合理使用”采取封闭式立法模式,若直接将 GenAI 语料训练的数据使用行为增列为第十三种法定情形,将违背封闭式的立法宗旨^[34]。然而,《著作权法》第二十四条还设有“法律、行政法规规定的其他情形”这一兜底条款,且《中华人民共和国著作权法实施条例》(以下简称《著作权法实施条例》)第二十一条对该条款作出进一步解释:使用可以不经著作权人许可的已经发表的作品,不得影响该作品的正常使用,也不得不合理地损害著作权人的合法权益^[35]。这一制度设计为 GenAI 语料训练的数据使用行为适用“合理使用”提供了法律依据。基于此,建议通过修订《著作权法实施条例》,针对 GenAI 语料训练数据使用行为的特点,补充完善相应的“合理使用”适用条件,在保护著作权人合法权益的同时,为 GenAI 技术创新发展提供必要的制度空间。

4.2.2 设定 GenAI “合理使用”的柔性边界

GenAI 语料训练的“合理使用”认定需审慎设定边界。欧盟和英国限制严格,仅允许“非商业用途”的 GenAI 语料训练行为,英国多次尝试通过修法扩展到“商业目的”,但遭到内容创作产业的反对。在界定 GenAI 语料训练的“合理使用”边界时,单纯以“商业目的”作为二分标准存在明显局限性。实际上,绝大多数 GenAI 训练行为都具有商

业属性,若将商业用途完全排除在合理使用范围之外,将导致制度的适用空间过于狭窄,架空了合理使用制度的功能价值,制约了技术创新;反之,若将商业用途全部纳入,则会导致利益的“天平”过度向人工智能企业倾斜,损害了著作权人的核心权利。相比之下,无论是日本的“非享受性使用”标准,还是美国的“四要素”和“转换性使用”的判断规则,均未将商业目的作为绝对排除要件,而是抓住该使用行为是否与著作权作品的潜在市场形成竞争关系,是否实质性损害权利人利益的底层逻辑来划定“边界”。这种基于市场替代效应的弹性判定标准,更能契合 GenAI 技术发展的特性。

因此,我国在构建限定条件时,应当摒弃简单的二分法,转而采用更具包容性和适应性的判定标准。建议我国借鉴日本和美国的制度设计,为 GenAI 语料训练限定一个柔性的“合理使用”边界。我国《著作权法实施条例》引入了“三步检验标准”,提出“合理使用”应当满足“不影响作品的正常使用”和“不会不合理地损害著作权人合法权益”两个条件^[36],为新的合理使用行为的认定创设了制度空间。然而,这一标准对 GenAI 语料训练行为的适用仍显宽泛。建议在现有框架下融入美国“转换性使用”规则的判断逻辑,若语料训练行为赋予原作品新的表达、意义或功能,且生成内容未对原作品市场构成实质性替代,则可认定为合理使用。

我国对于适当引入“转换性使用”规则具有法律基础,《著作权法实施条例》确立的“三步检验标准”为突破封闭式立法模式提供了制度接口,也为引入“转换性使用”规则创造了释法空间。这一制度演进在司法实践中得到强化,最高人民法院 2011 年发布的司法解释明确可以考虑将使用行为的性质与目的、作品性质、使用程度及市场影响等要素纳入合理使用认定标准,实质上为“转换性使用”规则的适用提供了司法依据^[37]。目前,该规则也在多个法院的裁判文书中得到实践应用。

总而言之,对于明确落在 12 种“合理使用”行为之外的 GenAI 语料训练的数据使用行为,可以在“三步检验标准”的基础上,进一步结合“转换性使用”的判断规则进行释法构建,从数据使用目的、特征以及是否会与原作品的市场构成潜在竞争而损

害著作权人利益来限定“柔性”边界，这更符合我国当前的 GenAI 发展现状和法律制度环境。

4.3 制定适用 GenAI 的知识共享协议方案，促进行业自律规范发展

无论是出台行政监管制度，还是通过修法和释法引入 GenAI “合理使用”法律制度，都需要时间。然而，GenAI 技术的发展不会因监管缺失而停滞。在明确的知识产权监管制度出台前，获取著作权人的许可仍是规避风险的主要方式。例如，OpenAI 与新闻集团达成多年期协议，支付超 2.5 亿美元，集中许可《华尔街日报》等 10 余家媒体的内容^[38]。这种集中许可模式可一次性满足海量数据需求，比逐一许可更节约成本。不过，集中许可仍需向众多著作权人支付费用，叠加之后仍会使 GenAI 公司难以承受。

以创作共享、使用共享为理念发展起来的知识共享协议 (Creative Commons License，简称“CC 协议”) 的许可模式，可能会成为克服 GenAI 语料训练著作权许可成本高昂难题的可行的行业合规方案。在 CC 协议框架下，著作权人可以围绕保留署名权、仅允许非商业性使用、禁止对作品进行演绎、基于作品的二次创作必须以相同方式共享这四项核心权利要求，制定 CC 许可格式协议，在允许他人免费使用作品的同时，保留一部分的著作权^[39]。CC 协议已在全球范围推行，并根据不同国家和地区的版权法律制度进行了兼容设计。CC 协议受到了全球内容创作者的普遍欢迎，共计有数百万个网站平台发布了超过 25 亿个采用 CC 协议的作品^[40]。

我国于 2006 年正式引入 CC 协议，并根据《著作权法》和相关法律法规，从作者署名权、邻接权、限制性条款等维度对协议进行了本土化修订，目前已经更新到 3.0 版本^[41]。不过，由于宣传性不足、与著作权法律制度的协调性不足、商业兼容性不足等原因，CC 协议仅在摄影、音乐作品和一些开放获取期刊上使用，尚待进一步推广。尽管如此，CC 协议作为一种弹性的许可授权模式，其双向选择特性与网络环境下作品传播模式高度契合，有效应对了传统单向授权机制与网络作品使用需求之间的结构性矛盾。学界普遍认为，CC 协议能够显著降低网络作品使用中的授权成本，解决著作权信息不明确

等现实问题，为网络作品的合法使用提供了有效解决路径^[42-43]。鉴于 GenAI 训练数据主要来源于网络作品，且开发者面临授权对象不明确、授权成本高昂等与传统网络作品使用相似的法律困境，CC 协议的授权模式与 GenAI 语料训练场景具有天然的适配性，可为 GenAI 的合规发展提供有效的解决路径。

制定并推广适用于 GenAI 语料训练的 CC 许可协议方案，以及基于 CC 协议建立提供训练数据的共享平台，能够在一定程度上解决现行监管制度存在的不确定性。通过 CC 协议标识，GenAI 开发者能够明确识别著作权人授权的数据使用范围，从而规避未经许可的数据使用。与《著作权法》“合理使用”制度的被动适用不同，CC 协议采用主动授权模式，要求著作权人在保留部分权利的前提下，明确许可 GenAI 开发者在特定条件下免费使用其作品。这种机制在“完全保留权利”和“放弃所有权利”之间开创了中间道路，赋予著作权人灵活设定使用条件的自主权。此外，CC 协议采用预设许可条款的共享模式，相较于传统的集中许可机制，不仅显著降低了 GenAI 企业的授权成本，还为其提供了明确且可预期的数据使用自由空间。

5 结 语

全球范围内，GenAI 语料训练引发的著作权侵权风险备受关注，美国、欧盟、英国和日本已采取不同程度的知识产权监管措施，而我国在此领域尚存监管空白。为兼顾 GenAI 企业与传统内容创作者的利益，我国应制定灵活的行政监管指导原则，并在著作权法律制度中完善适用 GenAI 语料训练的“合理使用”制度，“合理使用”制度应融合“三步检验标准”与“转换性使用”，从数据使用目的、特征及潜在竞争损害等方面限定柔性边界。同时，发展适用 GenAI 语料训练的知识共享协议方案，作为有效的行业合规路径。

尽管本研究致力于全面分析 GenAI 语料训练的域外知识产权监管制度，并据此提出符合我国国情的知识产权风险监管和合规方案，然而，GenAI 语料训练的知识产权风险目前仍处在发展初期，相关诉讼均在审理阶段，因此当前的研究无法考虑到 GenAI 语料训练的知识产权侵权的所有可能的情况，限制了监管制度研究的深度。展望未来，应基于更

多成熟的诉讼案例,对“合理使用”制度边界进行更细化的设计,同时研究如何将适用于 GenAI 语料训练的知识共享协议机制与现有法律制度进行协调,从而更好地应对 GenAI 技术更迭带来的新情况、新问题。

参 考 文 献

- [1] 人民网财经研究院,至顶科技.开启智能新时代:2024 年中国 AI 大模型产业发展报告[R].北京:人民网财经研究院,2024.
- [2] 张凌寒.加快建设人工智能大模型中文训练数据语料库[J].人民论坛·学术前沿,2024(13):57-71.
- [3] Pope A. NYT v. OpenAI: The Times's About-Face [EB/OL]. [2024-12-03]. <https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-times-about-face/>.
- [4] 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.
- [5] 李蔚容,吕君彦.论作品独创性的判断标准[J].Open Journal of Legal Science 法学,2024,12(4):2436-2439.
- [6] 蔡睿.人工智能训练语料的合理使用制度建构[J].荆楚法学,2024(5):91-101.
- [7] Stanford University. 2024 AI Index Report [R]. California: Stanford University, 2024.
- [8] The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence [EB/OL]. [2024-12-03]. <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>.
- [9] Congressional Research Service. Generative Artificial Intelligence and Copyright Law [EB/OL]. [2024-12-03]. <https://crsreports.congress.gov/product/pdf/LSB/LSB10922>.
- [10] U.S. Copyright Office. U. S. Copyright Office Fair Use Index [EB/OL]. [2024-12-03]. <https://www.copyright.gov/fair-use/>.
- [11] 熊琦,陈子懿.美国人工智能模型训练合理使用认定的成案经验研究[J].科技与法律,2024(6):11-23.
- [12] U.S. District Court for the District of Delaware. Thomson Reuters Enterprise Centre GmbH et al v. ROSS Intelligence Inc [EB/OL]. [2025-02-14]. https://www.ded.uscourts.gov/sites/ded/files/opinions/20-613_5.pdf.
- [13] 李国庆.谷歌图书馆案 The Authors Guild, Inc. v Google, Inc. 判决述评——以合理使用制度为视角[J].中国版权,2016(4):33-36.
- [14] European Parliament. EU Artificial Intelligence Act [EB/OL]. [2025-01-14]. <https://artificialintelligenceact.eu/recital/105/>.
- [15] Official Journal of the European Union. Directive(EU) 2019/790 [EB/OL]. [2024-12-03]. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0790>.
- [16] Browne R. SAP Boss Warns Against Regulating AI, Says Europe Risks Falling Behind U. S., China [EB/OL]. [2024-12-03]. <https://www.cnbc.com/2024/10/22/sap-ceo-urges-europe-not-to-regulate-ai-says-will-put-region-behind.html>.
- [17] 林小勇.未来媒体蓝皮书:中国未来媒体研究报告(2023)[M].北京:社会科学文献出版社,2023.
- [18] Davies C W. Getty Images v Stability AI: The Implications for UK Copyright Law and Licensing [EB/OL]. [2024-12-03]. <https://www.pinsentmasons.com/out-law/analysis/getty-images-v-stability-ai-implications-copyright-law-licensing>.
- [19] UK Gov. Copyright, Designs and Patents Act 1988 [EB/OL]. [2024-12-03]. <https://www.legislation.gov.uk/ukpga/1988/48/section/29A>.
- [20] UK Gov. Artificial Intelligence and Intellectual Property: Copyright and Patents; Government Response to Consultation [EB/OL]. [2025-01-14]. <https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents-government-response-to-consultation>.
- [21] Library Services. Copyright and TDM [EB/OL]. [2024-12-03]. <https://library.cranfield.ac.uk/text-and-data-mining/copyright>.
- [22] UK Gov. Copyright and Artificial Intelligence [EB/OL]. [2025-02-27]. <https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence/copyright-and-artificial-intelligence#our-proposed-approach>.
- [23] UK Gov. A Pro-innovation Approach to AI Regulation [EB/OL]. [2024-12-03]. <https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper>.
- [24] Nikkei Asia. Japan to Roll Out \$ 65bn in Support for Chips, AI [EB/OL]. [2025-01-14]. <https://asia.nikkei.com/Business/Tech/Semiconductors/Japan-to-roll-out-65bn-in-support-for-chips-AI>.
- [25] International Trade Administration. Japan Generative Artificial Intelligence [EB/OL]. [2024-12-03]. <https://www.trade.gov/market-intelligence/japan-generative-artificial-intelligence>.
- [26] Ministry of Internal Affairs and Communications Ministry of Economy, Trade and Industry. AI Guidelines for Business Ver1.0 [EB/OL]. [2025-02-27]. https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_9.pdf.
- [27] 中村合同特許法律事務所(株). The Copyright Act Revised in 2018 Will Further Improve the Machine Learning Environment in Japan [EB/OL]. [2024-12-03]. https://www.nakapat.gr.jp/ja/legal_updates_eng/the-copyright-act-revised-in-2018-will-further-improve-the-machine-learning-environment-in-japan/.
- [28] Japan Law Translation. 著作権法(明治三十二年法律第三十九号)[EB/OL]. [2024-12-03]. <https://www.japaneselawtranslation.go.jp/ja/laws/view/4207>.

- [29] Japan Copyright Office. “General Understanding on AI and Copyright in Japan” Overview [EB/OL]. [2024-12-03]. https://www.bunka.go.jp/english/policy/copyright/pdf/94055801_01.pdf.
- [30] Dermawan A. Text and Data Mining Exceptions in the Development of Generative AI Models: What the EU Member States Could Learn From the Japanese “Nonenjoyment” Purposes? [J]. The Journal of World Intellectual Property, 2024, 27 (1): 44-68.
- [31] 中央网络安全和信息化委员会办公室. 生成式人工智能服务管理暂行办法 [EB/OL]. [2024-12-03]. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.
- [32] 全国网络安全标准化技术委员会. 生成式人工智能服务安全基本要求 [EB/OL]. [2024-12-03]. <https://www.tc260.org.cn/upload/2024-03-01/1709282398070082466.pdf>.
- [33] 中华人民共和国中央人民政府. 中华人民共和国著作权法 [EB/OL]. [2024-12-03]. https://www.gov.cn/guoqing/2021-10/29/content_5647633.htm.
- [34] 张新宝, 卞龙. 生成式人工智能训练语料的著作权保护 [J]. 荆楚法学, 2024 (5): 77-90.
- [35] 中华人民共和国中央人民政府. 中华人民共和国著作权法实施条例 [EB/OL]. [2024-12-03]. https://www.gov.cn/gongbao/content/2011/content_1860739.htm.
- [36] 王思涵. 文本与数据挖掘著作权合理使用分析及其认定 [J]. 法学, 2023, 11 (6): 5130-5135.
- [37] 中华人民共和国最高人民法院公报. 最高人民法院印发《关于充分发挥知识产权审判职能作用推动社会主义文化大发展大繁荣和促进经济自主协调发展若干问题的意见》的通知 [EB/OL]. [2025-03-03]. <http://gongbao.court.gov.cn/Details/90857967c518c766c368851b1b705a.html>.
- [38] News Corp. News Corp and OpenAI Sign Landmark Multi-Year Global Partnership [EB/OL]. [2024-12-03]. <https://investors.newscorp.com/news-releases/news-release-details/news-corp-and-openai-sign-landmark-multi-year-global-partnership>.
- [39] Creative Commons. About CC Licenses [EB/OL]. [2024-12-03]. <https://creativecommons.org/share-your-work/cclicenses/>.
- [40] Creative Commons. Technology Platforms [EB/OL]. [2025-03-03]. <https://creativecommons.org/share-your-work/platform/>.
- [41] 知识共享中国大陆. 中国大陆版 CC 协议正式从 2.5 版本更新为 3.0 版本 [EB/OL]. [2025-03-03]. <https://creativecommons.net.cn/a-31/>.
- [42] 陈丽丽, 周小音. 知识共享协议的应用现状及措施分析 [J]. 法制博览, 2018 (21): 98-99.
- [43] 朱钰涛, 黄武双. 网络背景下知识共享协议在我国的适用 [J]. 人民司法, 2020 (7): 91-96, 101.

(责任编辑: 郭沫含)

(上接第 86 页)

- [24] Anderson W, Krathwohl R. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives: Complete Edition [M]. Addison Wesley Longman, Inc., 2001.
- [25] 刘玮, 陈锦博. 搜索即学习视角下知识捕获与学习效果的影响因素研究——以健康信息学习型搜索为例 [J]. 信息资源管理学报, 2023, 13 (6): 133-143, 155.
- [26] Urgo K, Arguello J, Capra R. The Effects of Learning Objectives on Searchers' Perceptions and Behaviors [C] //ICTIR'20: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval. ACM, 2020.
- [27] 沈旺, 杨真, 陈凤娟. 数字图书馆用户认知风格对信息检索行为影响研究 [J]. 情报理论与实践, 2014, 37 (8): 18-21, 31.
- [28] 孙晓宁, 姚青. 信息搜索用户学习行为投入影响研究: 基于认知风格与自我效能 [J]. 情报理论与实践, 2020, 43 (10): 99-107.
- [29] 秘塔 AI 搜索 [EB/OL]. [2024-12-10]. <https://metaso.cn/>.
- [30] 廖月含. 高中生场认知风格与英语写作策略的相关性研究 [D]. 桂林: 广西师范大学, 2023.
- [31] 方舟, 洪采菲, 秦妹言, 等. 自我调节学习情境下任务类型和学习认知策略对视频类搜索行为的影响 [J]. 图书情报工作, 2021, 65 (21): 84-96.
- [32] 董焕晴, 曹高辉, 虞沛. 生成式 AI 工具用户信息搜寻行为形成路径——以数字素养与人工智能素养为视角 [J/OL]. 图书馆论坛: 1-9 [2024-12-10]. <http://kns.cnki.net/kcms/detail/44.1306.G2.20241129.0846.002>.
- [33] Gadiraju U, Yu R, Dietze S, et al. Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web [C] //Proceedings of the 2018 Conference on Human Information Interaction & Retrieval. New York: ACM, 2018: 2-11.
- [34] Jiun-Yu W, Chen X. Using Time Pressure and Note-taking to Prevent Digital Distraction Behavior and Enhance Online Search Performance: Perspectives from the Load Theory of Attention and Cognitive Control [J]. Computers in Human Behavior, 2018, 88: 244-254.
- [35] 邹艳春. 认知风格: 影响学生学业成就的心理分析 [J]. 教育评论, 2002 (1): 34-36.
- [36] 王惠敏, 田新瑞. 生成式人工智能引发的信息过载风险及其对策 [J]. 科技管理研究, 2024, 44 (20): 149-156.
- [37] Fisher M, Goddu K, Keil C. Searching for Explanations: How the Internet Inflates Estimates of Internal Knowledge [J]. Journal of Experimental Psychology: General, 2015, 144 (3): 674-687.

(责任编辑: 郭沫含)