

数据产权“三权分置”视角下人工智能训练数据的治理困境与破解路径

郑泳君^{1*}

(¹ 华南农业大学 人文与法学学院, 广东 广州 510000)

摘要: 生成式人工智能的迅猛发展使其对高质量训练数据的需求呈指数级增长, 版权密集型、体系化的出版内容数据因此成为关键生产要素。然而, 传统以“作品”为核心的著作权保护框架与数据要素高频、大规模的流通需求发生严重冲突, 导致数据权属模糊、授权机制失灵、利益分配失衡等一系列治理挑战。本文以《关于构建数据基础制度更好发挥数据要素作用的意见》(“数据二十条”)提出的数据资源持有权、数据加工使用权、数据产品经营权“三权分置”的产权框架为理论核心, 剖析人工智能模型训练场景下出版数据治理的现实困境。研究认为, 单纯依赖版权的事后救济或合理使用的有限扩张均无法化解根本矛盾。出路在于构建以数据产权为核心的新型治理范式, 通过明确“三权”的法律内涵与行权边界, 并辅以配套机制, 方能在激励数据要素供给、保障各方主体权益与促进人工智能产业创新发展之间实现动态平衡。

关键词: 数据产权; 三权分置; 人工智能训练数据; 数据治理

DOI: <https://doi.org/10.71411/zgfx.2026.v1i1.862>

Data Property Rights "Tripartite Separation" Perspective on the Governance Dilemmas and Solutions for AI Training Data

Zheng Yongjun^{1*}

(¹ South China Agricultural University, School of Humanities and Law, Guangzhou, Guangdong, 510000, China)

Abstract: The rapid growth of generative AI has exponentially increased demand for high-quality training data, making copyright-intensive published content crucial. However, traditional copyright frameworks clash with data circulation needs, causing governance challenges like unclear ownership and licensing failures. Using China's "tripartite separation" data property rights framework (from the "Data Twenty Articles") as a theoretical basis, this study analyzes publishing data governance in AI training. It concludes that neither copyright remedies nor expanded fair use can resolve these issues, proposing instead a data property rights-centered governance model to balance data supply, rights protection, and AI innovation.

Keywords: Data property rights; Tripartite separation of rights; AI training data; Data governance

作者简介: 郑泳君 (2003-), 女, 广东广州, 学士, 研究方向: 知识产权、数字法治

通讯作者: 郑泳君, 通讯邮箱: 1606763549@qq.com

引言

人工智能技术正在改变社会生产的原有结构，它的发展离不开大规模、高质量的训练数据支撑。正如邢源恒指出，“出版数据作为高质量训练资源对提升模型的算法性能至关重要”^[1]，经过严格编审的出版内容，凭着自身的体系性和可靠性，成了大模型训练的“优质养料”。但数据的生产要素属性，和它在法律上的权属定位却形成了尖锐对立。一方面，科技公司从技术需求出发，更愿意主张模型训练是对数据的“合理使用”或“非表达性使用”，以此避开复杂的授权流程和成本；另一方面，出版机构却面临着“数据资产收益受损及技术累积导致的控制权弱化”的困境。

这种冲突的核心问题，在于传统版权制度的静态赋权方式，跟数据要素需要动态流通的需求不合拍。研究提到，新一轮科技革命让数据的存储、分析和互联能力呈指数级增长，但传统知识产权制度很难完全适配数据保护的需求，就连欧盟《数据法案》这类国际立法，也要面对和现有知识产权规则协调的难题，这说明制度失配是个普遍问题^[2]，其具体表现在“出版内容数据产权规范不明确、授权交易机制缺失、侵权判定困难、主体利益分配失衡”^[3]。现行《著作权法》对缺乏独创性的数据集合保护不足，《反不正当竞争法》虽然能提供行为规制，但总是事后补救、比较被动，无法给数据流通提供事前确权保障和稳定的交易预期。此外，现有基于传统信息产权保护的方案存在“保护不周延、不充分之处”，很难满足数据集合作为新型生产要素的治理需求^[4]。所以，寻求超越版权框架的治理方案，已然成了推动人工智能产业和出版产业协同发展的时代命题。

1 理论镜鉴：数据产权“三权分置”的框架与法理基础

为了解决数据要素的权属难题，国家层面出台“数据二十条”，确立了“数据资源持有权、数据加工使用权、数据产品经营权”分置的产权运行机制。这个框架给人工智能训练数据治理提供了全新的理论视角和实践路径，而且和国际主流做法、国内立法方向都能契合。

1.1 数据产权保护的法理正当性

数据产权的合理性，既有充分的法理支撑，也有实际实践基础。孙嘉宇从洛克的劳动财产理论出发，论证了数据产权保护的正当性——“数据从生产、流转应用到衍生凝结多方智力劳动与资源投入”^[5]，所以对付出实质性劳动的数据持有者赋予财产性权利，符合自然法的基本理念。出版机构把零散内容收集起来，经过清洗、标注、结构化处理，形成能用于模型训练的高质量数据集，这个过程投入的大量劳动和资本，正是新质生产力背景下“数据要素创新”的核心，自然该通过产权制度予以保护^[6]。此外，传统知识产权理论里的劳动价值说、个人自由主义，已经难以适应 AIGC 的发展需求，从数据要素视角界定产权，更能回应利益平衡的现实诉求，而“三权分置”正是这种视角的制度化体现。

1.2 “三权分置”框架的结构合理性

AI 训练数据的“采集——加工——商业化”全流程涉及出版机构、科技公司、交易平台等多元主体，通过拆分持有、使用、经营三权，恰好匹配该流程的主体分工。但数据产权的设计，不是为了追求对所有权的绝对控制，而是要让数据在流通和使用中释放价值。正如江波所指，数据治理应该“淡化数据所有权属性，强调数据使用权”^[7]。同时，“三权分置”作为数据要素创新性配置的核心制度，和人工智能立法的“创新法”定位高度契合——通过明确各主体的权利边界，既能保障数据供给，又能促进产业创新^[8]。而且，从知识产权到数据要素的视角转变，要求打破传统客体界定的局限，“三权分置”刚好适应了这种转变，允许数据在不同主体间根据具体场景高效流转。因此，相比把数据产权理解为所有权，数据财产权更合适，这能摆脱传统物权的思维

束缚，而“三权分置”正是这种财产权理念的灵活实践。

2 现实困境：人工智能训练数据治理的多维挑战

虽然“三权分置”在理论上给出了清晰路径，但放到人工智能模型训练的具体场景里，落地过程还面临不少挑战——这些挑战既来自技术本身的特性，也受限于制度的惯性。

2.1 权属模糊导致授权机制失灵

出版数据涉及作者、出版机构、技术平台等多个主体，权利链条十分复杂。深入分析“复合主体权属模糊”的问题，传统版权法以“作品单元”为权利边界，而数据要素化场景要求以“数据集”为治理单元，这种治理单元的错位，导致单个主体很难独立行使数据权利。而且现有知识产权制度对数据集保护存在空白，在数据库特殊权利这类保护路径，也存在不少争议。这进一步增加权属界定的难度。实践中，科技公司需要获取海量数据的授权，但传统的“事前授权”“单一主体授权”模式成本高、效率低，根本满足不了人工智能训练的需求，于是就引发了大规模的数据爬取与使用争议^[9]。一方面，著作权、商业秘密和反不正当竞争法对数据集的保护路径都有局限性；另一方面，AIGC的非表达性使用特征，让权利边界变得更模糊，给授权机制的设计增加了更多复杂度^[10]。

2.2 技术黑箱加剧侵权认定与维权困难

人工智能模型由于“算法黑箱”特性，使得数据的内部使用过程不透明。这种困境，在信息挖掘阶段，权利人较难证明训练方存在“接触”行为，到了信息处理阶段，模型算法拆解了“实质性相似”的判断空间。此外，AIGC的创作过程缺少人类智力劳动的直接参与，传统著作权侵权判定里的“独创性”“实质性相似”标准很难适用，这也与“非表达性使用”导致版权规则失效的观点能相互呼应^[11]。在 Robots.txt 协议，即便版权所有者采取技术手段选择退出，也因作品在互联网上存在多个副本而难以全面覆盖，且无法阻止从非法盗版网站抓取数据，这使得事后退出机制形同虚设。出版机构即使怀疑自己的数据被用于未经授权的训练，也常常因为没法证明“接触”和“实质性相似”，陷入维权无门的困境。除了医疗数据治理的研究^[12]，算法黑箱会导致风险难以追溯，在出版数据领域同样存在，且因数据规模更大，表现得更突出。

2.3 利益分配失衡抑制数据供给积极性

在数据要素的价值再分配中，出版机构往往处于弱势地位。研究揭示了“出版产业与人工智能产业的利益分配失衡问题”：科技公司靠着技术优势主导谈判，可能用“低价买断”的方式获取数据，而出版机构没法参与模型商业化后的持续分成^[3]。数据要素参与分配是新质生产力发展的必然要求，可当前的分配机制没能充分体现数据资源持有者的贡献，也可看出人工智能处理数据的权属不明导致激励不足。此外，这种利益失衡可能加剧“数据殖民主义”的风险，长期来看会挫伤出版机构进行数据化投入的积极性，最终可能导致高质量中文训练数据供给减少，损害整个人工智能产业的创新生态。

3 路径构建：以“三权分置”为核心的协同治理体系

要应对上述挑战，需把“数据二十条”的政策理念变为可操作的制度，搭建法律、技术与市场协同治理体系，充分吸收多领域的治理经验和立法智慧。

3.1 法律层面的制度创新

法律需细化“三权”的行使规则。这个过程要遵循“创新法”导向，又做好规范。基于前述

框架与法理,数据资源持有权应归属于对数据集投入实质性劳动的机构,其权能包括许可他人使用、转让权利和制止他人非法获取,机构媒体行使这项权利,从AIGC产业获得利润分红,正是“按数据要素分配”的具体结果^[11]。数据加工使用权是赋予科技公司的核心权利,但行使须以“合法授权”为前提,遵循“分层确权授权”思路^[12],根据数据的敏感程度设定不同的使用权限。数据产品经营权则允许对经过深度加工、形成独立价值的数据集进行市场化运营,这三项权利既可以分开配置,也可以集中在同一主体,形成灵活的权利安排。

3.2 市场机制的有效构建

搭建高效、可信的数据授权与交易机制。要解决海量数据授权的难题,就得积极探索数据产权集体管理模式。建议由国家级行业协会设立“数据产权管理中心”,统一负责费率谈判、批量授权与收益分配,并建立“许可费池”模式确保各方按贡献分享收益。同时,建立健全数据知识产权登记制度,制定统一的出版数据知识产权登记认定规范,并建立跨区域互认机制,如“行政机关登记—第三方机构实质审查”的登记模式^[13]。此外,还可以搭建多层次的数据交易所,形成成本低、效率高的场内交易环境,结合“三权分置”规则明确交易标的和权限,降低交易风险。

3.3 技术保障措施的强化

强化技术支持下的透明度和问责机制。要破解“算法黑箱”问题,就得要求科技公司做到算法透明、主动披露数据相关信息。“我国治理机制应适当扩展提供者履行数据透明义务的对象类型,将面向“监管端”的数据透明义务拓展至“用户端”^[14]。加工方可以定期公示数据来源、规模等摘要信息;建立分层分类的训练数据披露机制,对通用大模型强制要求披露版权作品的主要来源和授权类型。技术层面,出版机构可以用数字水印、区块链等技术提升数据溯源能力,科技公司则需要部署侵权过滤系统,履行“通知-取下”义务^[12]。此外,还可以借鉴的“技术合规保障制度”^[8],由监管部门、行业协会和技术专家共同组成审查委员会,以技术手段落实合规责任。

3.4 制度保障与跨领域协同

完善立法保障和跨领域协同治理。要破解当前人工智能训练数据治理困境,“三权分置”的政策框架转化为法律规范,可以通过行政法规、部门规章等形式细化数据产权规则,调整和丰富现有知识产权与竞争法制度。同时,针对当今“算法泛在”的表现,完善我国的算法备案制度,也是一剂良药,是“有效市场与有为政府相结合”原则在数字法治领域的延伸与创新^[15]。利益分配方面,要建立按贡献参与分配的机制,既保障出版机构的初始数据收益,也允许它们参与模型商业化后的持续分成。与此同时,要加强跨领域协同,推动出版行业和人工智能行业的标准对接,形成数据采集、加工、使用的行业规范,实现数据安全和创新效率的动态平衡。

4 结语

人工智能的发展浪潮不可阻挡,数据作为它的核心驱动力,地位也越发重要。但技术的跃进,不能以牺牲公平和秩序为代价。一直固守传统版权思维是解决不了人工智能训练数据带来的全新挑战。我们必须转向以数据产权为核心的新治理模式。数据集财产性权益的保护实践表明,数据处理者已经成为数字经济的关键参与者,权益设置要以激励数据处理行为为核心。而数据产权“三权分置”的改革构想,既符合中国数据资源丰富、应用场景多样的优势,也顺应了国际制度的演进趋势,为平衡数据流通效率和各方权益保障提供了可行的制度蓝图。通过法律层面明确产权边界、市场层面创新授权机制、技术层面保障透明可信、立法层面强化协同支撑,我们就能搭建起一个激励相容、可持续发展的人工智能数据生态,帮助解决出版数据的治理难题,为各类数据要素的合规流通提供参考,最终推动人工智能高质量发展,为新质生产力的发展注入持久动力!

参考文献:

- [1] 邢源恒. 模型训练场景下出版数据治理的挑战与优化路径[J]. 出版发行研究, 2025, (08): 13-20+12.
- [2] 单晓光. 数据知识产权中国方案的选择[J]. 人民论坛·学术前沿, 2023, (06): 38-47.
- [3] 叶悦. AI 大模型时代出版内容数据保护的理据与进路[J]. 出版与印刷, 2025, (01): 27-36.
- [4] 徐小奔. 人工智能时代数据集合财产性权益的侵权救济路径[J]. 法律适用, 2025, (06): 162-176.
- [5] 孙嘉宇. 数据产权: 生成式人工智能训练行为版权争议的规制路径[J]. 中国编辑, 2024, (08): 63-71.
- [6] 王雪琪. 数据要素赋能新质生产力“三新”要素的法律适用困境与出路[J]. 科技进步与对策, 2025, 42(07): 21-32.
- [7] 江波. 完善数据安全治理范式促进人工智能产业发展[J]. 信息安全研究, 2024, 10(08): 776-779.
- [8] 张敏. 新质生产力背景下人工智能立法的创新法定位研究[J]. 政法论丛, 2025, (03): 153-167.
- [9] 叶悦. AI 大模型时代出版内容数据保护的理据与进路[J]. 出版与印刷, 2025, (01): 27-36.
- [10] 季冬梅. 从知识产权到数据要素: AIGC 相关制度困境的化解[J]. 科技进步与对策, 2025, 42(02): 122-131.
- [11] 徐偲骅, 李凤. 数据资源持有权: 机构媒体应对 AI 侵权的升维因应[J]. 中国编辑, 2024, (11): 47-55.
- [12] 杨金铭, 王纳, 胡业勋, 等. 人工智能医疗中的法律风险防范[J]. 四川大学学报(医学版), 2025, 56(01): 143-148.
- [13] 李存正. 我国数据产权的保护困境及路径探索[J]. 传播与版权, 2025, (18): 102-106.
- [14] 张欣. 生成式人工智能的数据风险与治理路径[J]. 法律科学(西北政法大学学报), 2023, 41(05): 42-54.
- [15] 张吉豫. 论算法备案制度[J]. 东方法学, 2023, (02): 86-98.