农业图书情报学报
DOI：10.13998/j.cnki.issn1002-1248.24-0475

高质量AI数据体系面临的数据版权困境、应对策略解析与实施路径研究

张何灿[1]，易成岐[1*]，郭　鹏[2]，黄倩倩[1,3]，靳晓锟[4]

（1. 国家信息中心大数据发展部，北京 100045；2. 深圳数聚湾区大数据研究院战略研究中心，深圳 518048；3. 中国人民大学 信息资源管理学院，北京 100872；4. 中国科学院科技战略咨询研究院，北京 100190）

摘　要：[目的/意义]党的二十届三中全会决定明确提出，完善推动人工智能等战略性产业发展政策和治理体系。近年来，全球人工智能版权数据诉讼纷争频发，人工智能训练数据版权保护困境成为构建高质量AI数据体系面临的关键堵点和现实难题。[方法/过程]本研究在研究梳理人工智能数据版权保护相关学术研究和产业实践的基础上，系统性总结了应对数据版权困境的六大代表性做法，对比解析了不同做法的优缺点和适用性。[结果/结论]针对人工智能数据版权困境，即暂无既能促进人工智能版权数据供给又能兼顾数据版权保护工作的最优解问题，本研究在充分参考六大代表性做法解析和结合中国具备的四大独特优势基础上，研究提出系统妥善解决数据版权困境筑牢高质量AI数据体系的总体实施路径构想，分别为打造国家级人工智能数据版权一体化综合服务平台，探索推进适应人工智能发展的数据版权综合改革试点，建立完善人工智能数据版权相关立法并推动行业自律，以期对加大中国人工智能版权数据供给、制定相关政策和推动工作提供有益参考。

关键词：人工智能；AI数据体系；版权保护；数据版权；数据要素

中图分类号：TP3-05；TP271　　**文献标识码：**A　　**文章编号：**1002-1248（2024）09-0032-12

引用本文：张何灿，易成岐，郭鹏，等. 高质量AI数据体系面临的数据版权困境、应对策略解析与实施路径研究[J]. 农业图书情报学报，2024，36(9)：32-43.

0　引　言

党的二十届三中全会决定明确提出，完善推动人工智能等战略性产业发展政策和治理体系。近年来，随着以大模型为代表的人工智能技术快速发展演进，人工智能已成为中国发展新质生产力的重要核心引

收稿日期：2024-05-20

基金项目：国家自然科学基金专项项目“融合共票机制的元宇宙数字资产理论与方法研究”（62441206）；国家社会科学基金青年项目“面向多语种社会科学数据的线索发现方法研究”（22CTQ025）；国家社会科学基金青年项目“数据要素影响税收体系的机理及优化路径研究”（24CJY048）

作者简介：张何灿，研究实习员，国家信息中心大数据发展部人工智能处，研究方向为大数据与数字经济、人工智能语料体系等。郭鹏，高级咨询师，深圳数聚湾区大数据研究院（粤港澳大湾区大数据研究院）战略研究中心，研究方向为人工智能、软件工程。黄倩倩，助理研究员，国家信息中心大数据发展部人工智能处，中国人民大学信息资源管理学院，博士研究生，研究方向为人工智能、数据要素等。靳晓锟，博士研究生，中国科学院科技战略咨询研究院，研究方向为复杂网络、网络舆情等

*通信作者：易成岐，博士，副研究员，国家信息中心大数据发展部人工智能处，副处长，研究方向为人工智能、数据要素等。Email：yichengqi@sic.gov.cn

擎，成为中国抢占全球未来科技竞争制高点的必由之路，成为构建国家竞争新优势的关键驱动力[1]。如果说算力资源是人工智能大模型训练的“超级发动机”，那么AI数据体系则是人工智能大模型训练的“关键原材料”[2]，AI数据体系的规模和质量将直接决定大模型的精准理解能力、高效推理能力和场景适应能力。反之，如何构建高质量AI数据体系也成为当前人工智能发展的重要议题和瓶颈[3]。目前，中国努力推进数据要素高质量供给以及合规高效流通，数据工作体系基本形成、数据基础制度初步建立、数据供给力度不断增强、数据流通使用效能持续释放[4]，高质量AI数据体系建设初见成效。

与此同时，由于图书著作、期刊论文、版权音像制品等相关版权数据具有历史沿革长、知识密度高、训练价值大、法定事先授权等独特特征，高质量AI数据体系建设离不开引入大规模版权数据。然而，近年来人工智能企业对版权数据的渴求和版权所有者对数据版权保护之间的矛盾日益凸显，全球人工智能企业不当使用版权数据相关诉讼纷争频发。为此，各国纷纷尝试探索妥善解决人工智能训练数据版权保护困境的实施路径，如欧盟出台《人工智能法案》、韩国制定《生成式人工智能版权指南》、日本发布《人工智能与著作权相关问题的观点（草案）》等均对此问题有所提及。但总体而言，无论是侧重版权作品财产权的英国、美国等“版权体系”国家，还是以版权所有者人格权为重点的法国、德国等“作者权体系”国家，其现行的人工智能训练数据版权保护方式均尚不成熟，要么数据版权保护力度过于粗暴严格，或将严重阻滞人工智能版权数据供给，阻碍人工智能健康有序发展；要么数据版权保护力度过于放任宽松，或将遏制版权所有者原始创作创新动力。不难发现，全球尚无既能促进人工智能版权数据供给又能兼顾版权保护工作的最优解，人工智能训练数据版权保护困境（数据版权困境）成为构建高质量AI数据体系面临的关键堵点和现实难题。

在此背景下，更好兼顾人工智能版权数据高质量供给和数据版权妥善保护，对于促进高质量AI数据体系建设理论研究和实践创新、推动适应人工智能发展的数据版权制度改革、推进中国人工智能立法及进一步完善著作权法、推动中国人工智能和版权文化产业高质量协调发展等具有重要理论价值和现实意义。本研究在充分研究梳理人工智能数据版权保护相关学术研究与产业实践基础上，深度解析对比了全球重点企业应对人工智能训练数据版权保护困境的代表性做法及优缺点和适用性，最终研究提出应对数据版权困境筑牢高质量AI数据体系的总体实施路径构想，保障人工智能版权数据高质量供给同时兼顾数据版权妥善保护，以期对加大中国人工智能版权数据供给、制定相关政策和推动工作提供有益参考。

1　研究综述

长久以来，大部分人工智能模型只能通过有标注的数据进行模型训练，大模型问世之后无标注数据也可被大规模用于人工智能预训练[5]，进一步扩展了高质量AI数据体系构成，对高质量AI数据体系建设提出了更高要求。2022年，《中共中央 国务院关于构建数据基础制度更好发挥数据要素作用的意见（中发〔2022〕32号）》根据数据生成来源和数据价值属性，将数据分为公共数据、企业数据、个人信息数据三大类[6]，无论是公共数据还是企业数据和个人信息数据，对提升大模型面向多任务需求时所表现出的通用推理能力均十分重要，但与此同时，在公共数据、企业数据、个人信息数据形成规模化数据集构建过程中，也会不可避免地收集、涵盖、涉及具有版权属性的语料数据，并且由于版权数据的高质量、高价值、高可用等特点，现已经成为大模型能力制胜的关键因素[7]，当然，由此引发大模型企业对于版权数据的极度渴求和版权所有者对版权作品的有效保护之间难以兼顾的矛盾且愈演愈烈，人工智能训练数据版权保护困境由此再次成为产学界关注热点。

1.1　人工智能训练数据版权问题成因研究

业界认为主要有数据需求海量性与多样性、“事

DOI：10.13998/j.cnki.issn1002-1248.24-0475

前授权”式使用付费模式高本低效、合理使用难适用三大因素引发人工智能训练数据版权保护问题。在数据需求海量性与多样性方面，人工智能大模型对包括版权数据在内的训练数据规模和质量要求显著提升，其训练数据集涉及文本、图片、视频、音频、代码等多种类多模态，但大部分企业选择通过爬虫、API接口对接等自动化手段从互联网上直接复制获取，涉及大量达到“独创性”标准的版权作品，势必造成著作权侵犯[8]。在“事前授权”使用付费模式高本低效方面，有学者认为“事前授权”式使用付费模式易导致数据交易流程冗杂及交易效率低下、数据获取识别成本和数据交易谈判成本高等问题，从而加重了人工智能企业的版权负担，不契合人工智能数据训练需求[9]。在合理使用难适用方面，中国现行《著作权法》第二十四条第一款规定了合理使用情形，其中的第（一）项“个人学习、研究”、第（二）项“适当引用”、第（六）项“科学研究、第（十三）项其他情形”涉及人工智能，学界以第（一）项的自然人主体和非营利性要求、第（二）项的“介绍、评论和说明”要求、第（六）项少量和科研目的要求、第（十三）项司法适用空间有限等考虑出发，对于现行合理使用情形难以适用人工智能产业实际现状进行阐述[10-12]，也有学者对合理使用的认定方法提出一定质疑，认为人工智能创作过程的各个阶段使用作品均要受到作者复制权约束，传统的“三步检验法”难以继续适用人工智能[13]。

1.2 人工智能训练数据版权问题现状研究

针对人工智能训练阶段，学术界以中国现行《著作权法》为重点研究对象，从立法原意、原则导向等角度出发，认为部分法律条文在人工智能训练数据集版权客体认定、人工智能侵权使用版权数据情形认定、人工智能侵犯的版权细分权利认定、人工智能侵犯使用版权归责原则适用等关键问题界定不明确或不合理，对明确认定上述问题和合理解释进行了深入探讨。其中，对于人工智能训练数据集版权客体认定研究，部分学者认为训练数据集有正当和必要作为单独的版权客体予以保护，具体适用上主要提出通过增设邻接权[14]、增设新型知识产权[15]、申请涉及数据的专利权保护[16]、调整《反不正当竞争法》[17]、作为商业秘密[18]、制定数据权益保护的专门立法[19]等主张。对于人工智能侵权使用版权数据情形认定研究，部分学者认为由于用户难以察觉和举证，传统的“接触+相似”侵权认定方法难以招架人工智能场景[9]。对于侵犯的版权细分权利认定研究，学界大体认为人工智能训练阶段涉及复制权、改编权、翻译权等版权细分权利，对复制权的侵犯大部分学者表示赞同，认为只要涉及保护期内版权作品的扫描、拷贝、提取等行为都应纳入复制权的控制范畴[20,21]，少数表示反对，认为机器学习各个阶段中的数据处理行为仅构成对版权作品内容的“非作品性使用”，因此并不构成版权侵权[22]；对侵犯改编权的判定有观点认为无论何种处理都可能落入改编权的范畴[7]，也有学者指出应取决于人工智能将原作品分解和重新组合的形式与程度[23]。对侵犯翻译权的判定有学者认为，将普通文本翻译为计算机可读文本不构成侵权，但《著作权法》意义上将作品翻译为其他人类语言的行为涉及[24]。对于人工智能侵犯使用版权归责原则适用研究，学界对于适用过错责任原则或是无过错责任原则存在分歧争议，前者出于人工智能行为主体复杂、版权侵权行为风险较为可控等考虑，主张适用过错责任原则[25,26]；后者制度成本更低、版权权益救济及时充分等考虑，主张适用无过错责任原则[27,28]更加严格追责。

1.3 人工智能训练数据版权问题实证研究

随着全球范围内人工智能版权数据诉讼案件数量增加，学界对于司法实践中真实案例也开始关注研究，如有学者重点针对侵权行为、价值考量、业界影响对“AI文生图”案进行了剖析解读，认为“AI文生图”属于美术作品其著作权归属于利用人工智能生成图片的人[29]。然而，现有研究更多聚焦于微观视角下的人工智能生成阶段（人工智能生成物版权性）的版权保护案例研究。据此，本研究基于互联网公开报道不完全统计，2023年以来至2024年9月底全球人工智

能训练数据版权保护相关诉讼高达28起，集中发生在中国与美国，分别为3起、25起。美国方面，诉讼提交地横跨多个司法管辖区，包括特拉华州、纽约南部、加州北部及田纳西州中部等，版权所有者主要为新闻机构、出版社、图片社、音乐出版商、书籍作者等角色，人工智能企业中以OpenAI（9起）和微软公司（5起）被起诉次数最多，Stability AI、Meta、GitHub、Alphabet、英伟达等人工智能相关企业也有牵扯。判决结果上，美国人工智能数据版权保护相关诉讼案目前均在审理阶段，尚未最终判决。中国方面，版权所有者角色、诉讼理由情况与美国基本一致，诉讼提交地集中在北京和广州。不同的是，中国已有案件宣布司法判决，即“AI生成奥特曼案”，明确了版权所有者对原创作品的合法权益，为全球人工智能版权保护工作提供了参考借鉴。

不同于高质量AI数据体系建设中的其他堵点难题，人工智能训练数据版权保护困境涉及法律、技术、产业、市场、政策、知识产权等多个方面，不能单靠某个单一途径解决全局性问题。目前，相关研究大多聚焦于各自的研究视角和判别方向，对于更好统筹构建综合技术、机制、政策等全局性解决框架或解决方案的思考略有不足。此外，学术界在部分原则性关键问题上还存在一定分歧争论且僵持不下，如人工智能训练数据集版权客体认定、人工智能侵犯使用版权作品归责适用原则等。再者，大部分研究停留在学术理论层面，尚缺乏对产业界实际做法的梳理分析和融合利用。

2　应对数据版权困境的六大代表性做法解析

正如OpenAI向美国政府“限制使用版权数据就是扼杀大模型发展”所述，全球人工智能企业并未因数据版权困境尚未得到有效解决而止步不前，而是不惜代价、想尽办法、多措并举对版权数据“应采尽采、应用尽用”，以期进一步提升大模型能力水平。当然，目前部分人工智能企业为了应对人工智能训练数据版权保护困境，也涌现出了6种不同代表性做法。

2.1　代表性做法1：人工智能企业与版权所有者签订许可使用合同，业界应用最为广泛且成熟的版权数据使用方式

（1）做法释义。人工智能企业与版权所有者签订许可使用合同，是指双方通过签订协议等法律文件获取版权所有者持有的版权数据授权，常见的合作形式包括版权数据一次性买断或转让、按期限有偿授权使用版权数据、按版权数据产生的收益反馈提成、人工智能企业为其提供技术服务等。人工智能企业与版权所有者签订许可使用合同，是目前业界最常见、最成熟、最广泛的版权数据使用方式。

（2）代表性案例。由于图片和音视频等模态数据训练存在数据开销大、采购成本高等因素，不少人工智能企业均有采取签订许可使用合同获取版权数据的公开报道记录，国内如华为云与中文在线、中广天择与万兴科技等，国外如OpenAI与《金融时报》、谷歌与Reddit等，从数量上看OpenAI和谷歌公司应用此方式最为广泛，合作次数分别达4次和3次（表1）。

2.2　代表性做法2：发起数据版权专项合作计划或组建联盟，尚在初期宣传探索性阶段

（1）做法释义。发起数据版权专项合作计划或组建联盟，是指多个人工智能企业或版权所有者通过发起专项计划、组建版权数据联盟、打造版权数据应用共同体等方式形成风险与利益共享的多方合作关系，以期实现对版权数据资源、人工智能技术、版权产生收益等方面的资源互补和互利共赢。

（2）代表性案例。一是Perplexity AI公司在被亚马逊等公司指控过度抓取数据后，于2024年7月推出“出版商计划”，该计划提出Perplexity AI在引入合作伙伴版权数据的同时，需要向合作伙伴反哺一定程度的收入分成及提供必要的技术支持。截至目前，《时代》《明镜周刊》《财富》《企业家》《德克萨斯论坛报》、WordPress.com等媒体或企业已成为“出版商计划”的首批合作伙伴。二是美国音乐数据集提供商

DOI：10.13998/j.cnki.issn1002-1248.24-0475

表1　部分人工智能企业与版权所有者开展商业合作情况

Table1　Commercial cooperation between some AI enterprises and copyright owners

达成合作时间	AI企业	版权所有者/著作权人	版权数据类型、协议期限及金额
2024年7月	微软	泰勒·弗朗西斯（Taylor & Francis）	论文期刊数据，协议期限不详、协议金额1 000万美元
2024年5月	OpenAI	美国新闻集团（News Corporation）	新闻数据，协议期限5年、协议金额超2.5亿美元
2024年4月	OpenAI	英国金融时报	新闻数据，协议期限金额不详
2024年2月	谷歌	Reddit平台	社交媒体数据，协议期限不详、协议金额6 000万美元
2024年1月	万兴科技	中广天择	视频数据，协议期限金额不详
2023年12月	OpenAI	施普林格出版集团（Axel Springer）	新闻数据，协议期限金额不详
2023年11月	谷歌	加拿大新闻出版商	新闻数据，协议期限不详、协议金额1亿加元（约合7 360万美元）
2023年10月	谷歌	德国Corint Media组织	新闻数据，协议期限不详、协议金额320万欧元（约合338万美元）
2023年9月	华为云	中文在线	包括文字音视频等文字数据，协议期限金额不详
2023年7月	OpenAI	美联社	新闻数据，协议期限金额不详

Rightsify、Rightsify旗下全球版权交易所Global Copyright Exchange（GCX）、美国图像授权服务商vAIsual、日本图片库提供商Pixta、德国数据市场Datarade、美国创意社区平台ado、美国视频和文本数据集提供商Calliope Networks等7家机构，于2024年6月共同发起成立“数据集提供商联盟（DPA）”，旨在建立推广符合各国现行著作权法精神的人工智能版权数据开发利用标准和法律框架，积极推动促进人工智能企业和版权数据提供商合作。目前，“出版商计划”和“数据集提供商联盟”皆只运行数月，尚处于初期宣传探索阶段，未披露取得的实质性进展或合作成果。

2.3　代表性做法3：人工智能企业引入版权声明机制，根据用户实际反馈从训练AI数据体系中撤销相关版权作品

（1）做法释义。用户声明机制主要包括选择退出及选择进入两种方式，前者原则上视版权所有者默认同意使用其版权作品，但是版权所有者有权告知并要求人工智能企业采取措施停止使用，后者则反之，即原则上视版权所有者默认不同意利用其版权作品，但版权所有者通过公开声明明确表示许可后，人工智能企业方可使用其版权作品用作模型训练。

（2）代表性案例。从互联网公开新闻报道和企业披露中，目前仅发现Stability AI公司与OpenAI宣布引入版权声明机制，且均选择退出机制作为具体声明方式。2023年3月，Stability AI公司宣布Stable Diffusion 3.0模型将允许版权所有者从合作公司Spawning的“Have I Been Trained”平台上查询，确认其版权作品是否被用于模型训练，并会根据版权所有者意愿从训练数据集中移除具有争议的版权数据。2024年5月，OpenAI公司宣布开发“媒体管理器（Media Manager）”工具，旨在帮助版权所有者识别其版权作品是否被用于人工智能大模型训练，若有涉及则允许版权所有者“选择是否退出”，并计划于2025年前正式投入使用。

2.4　代表性做法4：人工智能企业引入版权风险担保机制，有效缓解用户使用人工智能产品的版权保护问题担忧

（1）做法释义。人工智能企业为有效缓解人工智能产品用户对版权保护问题的担忧，在满足一定条件情况下，将会为使用其产品而造成版权赔偿的用户提供法律协助和资金担保援助。

（2）代表性案例。谷歌、微软、亚马逊、Getty Images、Shutterstock、Adobe、OpenAI等全球知名人工智能企业均曾明确作出过版权风险担保承诺。如2023年6月，Adobe公司正式推出面向企业客户的

Firefly产品，并保证使用过程中如果涉及版权侵权诉讼，相关费用均由Adobe公司承担赔偿。2023年9月，微软承诺只要在启动内置的安全防护和内容过滤器前提下，用户使用Copilot产品引起的版权侵权诉讼问题，均可由微软提供辩护并支付赔偿金。2023年11月，OpenAI宣布推出“版权盾”，承诺为面临版权索赔的ChatGPT用户辩护并承担相应法律费用。

2.5　代表性做法5：人工智能企业改用合成数据代替部分版权数据，成为业界较为通行的模型训练技术性做法

（1）做法释义。人工智能企业改用合成数据代替部分版权数据进行模型训练，其中，合成数据指通过人工智能特定技术对原始训练数据进行学习，从而生成具有相似特征的新数据并用于大模型训练，合成数据在基础大模型训练以及自动驾驶、金融、电信、医疗等领域行业或垂直大模型训练得到广泛应用。

（2）代表性案例。得益于在重点行业和关键场景的成功应用实践，加之人工智能合成物版权性法律尚未明晰，合成数据较少引发版权争议，在替代版权数据进行模型训练方面逐渐崭露头角，国内外商汤科技、Anthropic、OpenAI、英伟达等公司均采用大量合成数据进行模型训练。如2024年5月，商汤科技发布“日日新5.0”通用大模型，该模型采用混合专家架构，并声称利用大量合成数据进行模型训练。2024年3月，Anthropic公司发布Claude 3系列通用大模型，从公布的技术报告发现，Claude 3系列模型使用了合成数据。2024年3月，英伟达高级科学家Jim Fan推测，OpenAI公司在训练Sora过程中，大规模使用了游戏引擎（UE5）的生成数据。2024年6月，英伟达发布了Nemotron-4340B开源模型，该模型可以专为训练大语言模型生成合成数据。

2.6　代表性做法6：研发针对大模型的版权数据检测工具，便于人工智能企业自查和版权所有者维权使用

（1）做法释义。随着图像水印、音频识别和文本分析等技术的快速发展，针对大模型的版权数据检测工具应运而生，受到版权所有者的广泛关注。针对大模型的版权数据检测工具能够辅助人工智能企业进行自查，帮助版权所有者准确识别并快速存证数据版权侵权行为，能够进一步保障版权所有者的合法权益。

（2）代表性案例。版权数据检测工具在传统场景应用广泛且相对成熟，但针对大模型的版权数据检测工具目前仍处于初级发展阶段。2022年9月，Spawning公司基于LAION-5B开源数据集推出“Have I Been Trained”平台，可以帮助用户发现其版权照片是否被用于“文生图”模型训练。2024年3月，人工智能模型评估公司Patronus AI推出版权检测工具CopyrightCatcher，可以检测大模型的输出结果中是否含有侵权内容。2024年8月，初创公司ProRata宣称将推出一个结合聊天机器人和搜索引擎的平台，能够利用其专有算法识别发现人工智能模型使用的版权作品。

2.7　六大代表性做法对比分析

上述应对人工智能训练数据版权保护困境的六大代表性做法具有不同的优缺点和适用情形（表2）。具体来看：①两方签订许可使用合同方式获取版权数据效率最高、风险最低、适用范围最广，但存在版权数据采购议价成本高、批量获取个人持有版权数据效率偏低等局限，因此更适用于资金储备较为雄厚的人工智能企业，或是对数据质量规模和权威性有较高要求的科研院所、咨询机构等单位采用；②发起专项计划或组建联盟方式继承了签订合同的部分优点，并能够一定程度缓解版权数据采购议价成本较高的问题，但仍然面临暂未取得实质进展或成效、多方共识难达成、执行效率和灵活性不足等新挑战，因此适用于在业内具有一定影响力、话语权较大或版权数据资源独特等企业发起或参与；③引入版权声明机制直接省去了获取版权授权的溯源和采购成本，适用范围较为广泛，但也面临声明易被忽视、“退出”操作技术性要求较高、大量作品“退出”将影响大模型性能等挑战，多适用于有一定合规版权数据储备和技术能力的

DOI：10.13998/j.cnki.issn1002-1248.24-0475

人工智能企业；④引入版权风险担保机制可以提升企业口碑、增加社会信任，从而在一定程度减少人工智能用户和版权所有者之间的诉讼纷争，但其本质是将部分用户使用过程中触发的侵权责任转移至企业自身，并且保障条款往往对担保情形有一定额外要求，因此多适用于有一定合规版权数据储备或是法律资金资源充足的人工智能企业；⑤合成数据可以在一定程度上规避训练数据版权保护要求，具有生产数据效率高、成本低、可持续等优点，但如果合成数据是由真实版权数据生成而来，或将无法完全根除版权保护风险隐患，并且进一步加大侵犯版权察觉溯源取证的难度，此外，在部分场景下合成数据的知识密度与原始数据存在一定差距，若不当使用将导致“模型崩溃”，因此合成数据更适用于一些特定如数据原创性要求相对较低、版权数据规模要求相对较小等的场景应用；⑥版权数据检测工具能够缓解版权所有者察觉侵权和侵权取证维权难问题，提升版权作品的创作动力和创作环境，帮助人工智能企业提前发现未获授权的版权数据并予以应对，降低侵权相关诉讼纠纷，但当前适用于人工智能大模型的版权检测工具较少问世且技术本身亦不成熟，且频繁进行版权检测也会提高企业版权数据管理成本。

表2　应对数据版权困境的六大代表性做法解析

Table2　Analysis of six representative approaches to address the copyright data dilemma

代表性做法	优点	不足	侵权风险	适用情形
双方签订许可使用合同	获取版权数据效率最高、风险最低、适用范围最广	版权数据采购议价成本高、批量获取个人持有版权数据效率偏低	无	资金储备较为雄厚的人工智能企业，对数据质量规模和权威性有较高要求的科研院所、咨询机构等单位
发起专项计划或组建联盟	继承了签订许可使用合同的部分优点，一定程度缓解版权数据采购议价成本较高的问题	暂未取得实质进展或成效、多方共识难达成、执行效率和灵活性不足等	低	业内具有一定影响力、话语权较大或版权数据资源独特等企业发起或参与
引入版权声明机制	无获取版权授权的溯源和采购成本、适用范围广	声明易被忽视、操作技术性要求较高、大量作品“退出”将影响大模型性能等	中	有一定合规版权数据储备和技术能力的人工智能企业
引入版权风险担保机制	提升企业口碑、增加社会信任，在一定程度减少人工智能用户和版权所有者之间的诉讼纷争	部分用户使用过程中触发的侵权责任转移至企业自身，保障条款往往对担保情形有一定额外要求	中	有一定合规版权数据储备或是法律资金资源充足的人工智能企业
改用合成数据代替	生产数据效率高、成本低、可持续	无法完全根除版权保护风险隐患，进一步加大侵犯版权察觉溯源取证的难度	低	一些特定如数据原创性要求相对较低、版权数据规模要求相对较小等的场景应用
应用针对大模型的版权检测工具	缓解版权所有者察觉侵权和侵权取证维权问题，提升版权作品的创作动力和创作环境，帮助人工智能企业提前发现未获授权的版权数据	当前适用于人工智能大模型的监测工具较少问世且技术尚不成熟、提高企业版权数据管理成本	不涉及	具有公信力的第三方机构

3　应对数据版权困境筑牢高质量AI数据体系的总体实施路径

高质量AI数据体系面临的训练数据版权保护困境涉及法律、技术、市场、政策等多个方面，往往牵一发而动全身，不能简单粗暴解决、一蹴而就推动。在充分参考借鉴上述六大应对数据版权困境的代表性做法基础上，结合中国具备的新型举国体制、社会主义市场经济体制、超大规模市场、数据基础制度等方面独特优势，研究提出中国应对数据版权困境筑牢高质量AI数据体系的总体实施路径构想，具体如图1所示。

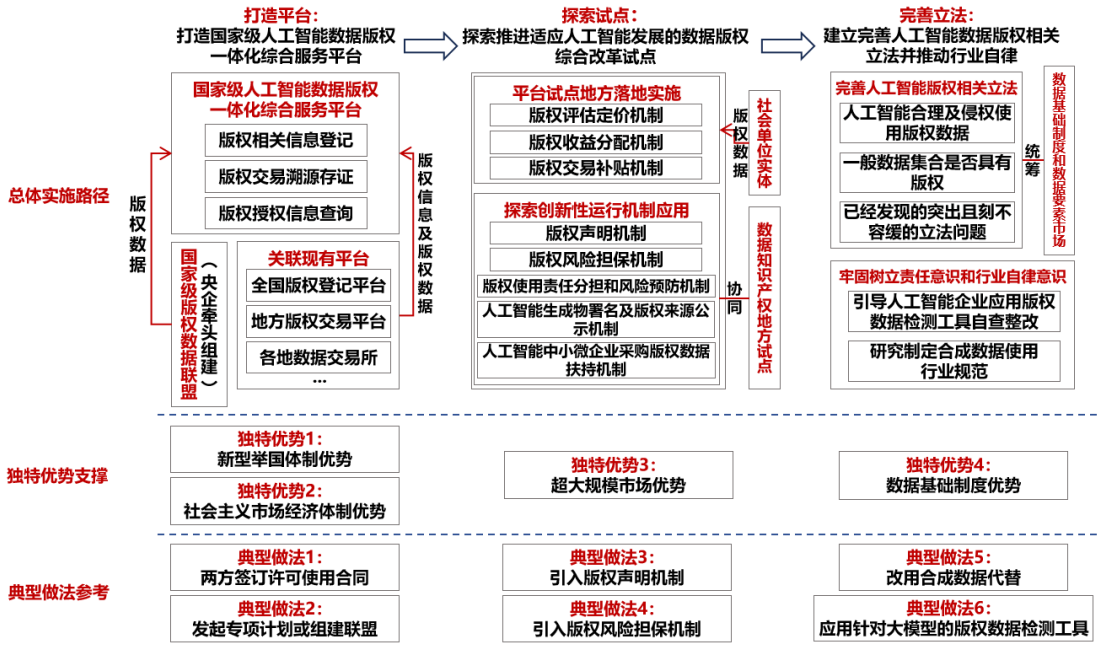

图1　应对数据版权困境筑牢高质量AI数据体系的总体实施路径

Fig.1　General implementation path to build a high-quality data system for AI to address the copyright data dilemma

3.1　打造平台：打造国家级人工智能数据版权一体化综合服务平台

不难发现，应对人工智能训练数据版权保护困境的六大代表性做法中，双方签订许可使用合同是当前最常见、最成熟、最广泛、风险最低的企业应对方式，加之完善立法方面可能存在一定程度的滞后性，因此当前应对数据版权困境的主要矛盾，应是进一步加强人工智能企业与版权所有者的合作共赢而非急于推动完善立法。具体举措上，首先，应充分发挥中国新型举国体制优势，科学谋划、统筹布局搭建国家级人工智能数据版权一体化综合服务平台。以国家级人工智能数据版权一体化综合服务平台的公信力和引导力为抓手，最大程度撮合增进人工智能企业与版权所有者版权的合作意愿。此外，国家级人工智能数据版权一体化综合服务平台应全面对接全国版权登记平台、地方版权交易平台及各地数据交易所等现有相关平台，集成上述平台的版权信息及版权数据资源，重点实现版权相关信息登记（包括版权作品、版权所有者、版权使用方等）、版权交易溯源存证（包括版权数据流转、交易、溯源、存证等）、版权授权信息查询等重要功能。其次，考虑到大量图书专著、期刊论文、新闻作品、电视节目等版权数据由央国企或行政事业单位掌握，应参考现有“发起专项计划或组建联盟”做法，充分发挥中国高水平社会主义市场经济体制优势，由央企牵头组建国家级版权数据联盟，在确保安全的前提下，整合共享各领域高质量版权数据资源，引接至国家级人工智能数据版权一体化综合服务平台，大幅提高全社会版权数据溯源效率和降低版权数据采购综合议价成本，为推动解决人工智能训练数据版权保护困境奠定服务平台基础和数据资源基础。

3.2　探索试点：探索推进适应人工智能发展的数据版权综合改革试点

为积极配合有效推动国家级人工智能数据版权一体化综合服务平台的长效可持续运行，应参考产业界的“引入版权声明机制”和“引入版权风险担保机制”的机制层面应用做法，充分发挥中国超大规模市场优势，探索推进适应人工智能发展的数据版权综合改革试点。一方面，推动国家级人工智能数据版权一

DOI：10.13998/j.cnki.issn1002-1248.24-0475

体化综合服务平台在具有条件的试点地方落地实施，在引导中央层面加大版权数据供给的基础上，积极鼓励社会各领域单位实体将掌握的版权数据对接至国家级人工智能数据版权一体化综合服务平台，同步试点建设版权评估定价机制、版权收益分配机制、版权交易补贴机制等，在促进版权所有者共享人工智能发展收益的基础上降低相关风险；另一方面，协同数据知识产权地方试点，积极探索版权声明机制、版权风险担保机制、版权使用责任分担和风险预防机制、人工智能生成物署名及版权来源公示机制、人工智能中小微企业采购版权数据扶持机制等创新性运行机制，在国家级人工智能数据版权一体化综合服务平台上的应用实践，根据应用效果逐步完善平台功能和机制设计，在持续加大人工智能企业和版权所有者合作力度和合作意愿的同时，更好兼顾人工智能发展和训练数据版权保护应对。

3.3 完善立法：建立完善人工智能数据版权相关立法并推动行业自律

首先，应辩证看待围绕人工智能数据版权问题推动立法的必要性：一方面，以大模型为代表的人工智能对现行版权相关法律体系造成了新的挑战；另一方面，法律的适用具有一定灵活性和弹性，技术的推陈出新不必然导致法律的修订或新设[29]。因此，在法律法规层面，不宜盲目扩大《著作权法》的合理使用，如将人工智能企业直接使用版权数据训练情形认定免责等，此举不仅不利于中国版权产业创新发展，还可能导致中国版权数据被境外轻易合法获取，严重影响中国人工智能全球博弈竞争。为此，应充分结合适应人工智能发展的数据版权综合改革试点成效，重点围绕人工智能合理及侵权使用版权数据、一般数据集合是否具有版权等原则性或关键性问题进一步研究界定，进一步梳理总结人工智能数据版权已经发现的突出且刻不容缓的法律问题，统筹数据基础制度和数据要素市场以发挥数据基础制度优势，适时启动并建立完善人工智能数据版权相关立法工作。此外，应参考产业界“改用合成数据代替”和“应用针对大模型的版权数据检测工具”做法，引导人工智能企业应用版权数据检测工具自查整改，研究制定合成数据使用行业规范，推动人工智能企业和相关机构牢固树立企业的责任意识和行业自律意识。

4 总结与展望

为解决促进人工智能版权数据供给和数据版权保护工作无法兼顾的人工智能训练数据版权保护困境，本研究在充分梳理人工智能数据版权保护相关学术研究和产业实践的基础上，对比解析了应对数据版权困境的六大代表性做法及相关优缺点和适用性，研究提出应对数据版权困境筑牢高质量AI数据体系的总体实施路径构想。下一步，拟持续跟踪并及时总结各国人工智能数据版权相关政策动向，进一步细化国家级人工智能数据版权一体化综合服务平台的功能设计，研究梳理适应人工智能发展的数据版权综合改革试点关键要点，强化人工智能数据版权顶层设计，为解决高质量AI数据体系面临的数据版权困境提供研究参考。

参考文献：

[1] 于凤霞. 抓住人工智能“牛鼻子”加快形成新质生产力[EB/OL]. (2024-01-10) [2024-05-11]. https://www. ndrc. gov. cn/wsdwhfz/202401/t20240110_1363194.html.

[2] 张文娟, 邓辉, 艾政阳, 等. 我国AI大模型数据集建设发展刍议[J]. 人工智能, 2024, 11(3): 85-95.
ZHANG W J, DENG H, AI Z Y, et al. On the construction and development of AI large model dataset in China[J]. AI-View, 2024, 11(3): 85-95.

[3] 腾讯研究院. AIGC发展趋势报告2023: 迎接人工智能的下一个时代[R/OL]. 北京: 腾讯研究院, 2023. https://research. tencent. com/report?id=AJJ.

[4] 盘和林, 茹少峰, 易成岐. 深入推进数字经济创新发展[N]. 经济日报, 2024-06-12(010).

[5] 蔡津津. AIGC时代新闻舆论工作新阵地——面向大模型的可信训练数据集与服务能力建设[J]. 中国传媒科技, 2023(10): 79-83.
CAI J J. A new position of news and public opinion work in AIGC

era - Credible training data set and service capacity building for large model[J]. Media science and technology of China, 2023(10): 79-83.

[6] 新华社. 中共中央 国务院关于构建数据基础制度更好发挥数据要素作用的意见[EB/OL]. (2022-12-19)[2024-05-11]. https://www.gov.cn/zhengce/2022-12/19/content_5732695.htm.

[7] 高雅文, 来小鹏. 生成式人工智能语料版权问题研究[J]. 出版广角, 2024(5): 27-34.
GAO Y W, LAI X P. Research on copyright of generative artificial intelligence corpus[J]. View on publishing, 2024(5): 27-34.

[8] 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.
ZHANG T. Legal risks of generative AI training datasets and inclusive prudential regulation[J]. Journal of comparative law, 2024(4): 86-103.

[9] 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.
ZHANG P. The obstacles and solutions of copyright system in artificial intelligence content generation mechanism[J]. Science of law (Journal of northwest university of political science and law), 2024, 42(3): 18-31.

[10] 周文康, 费艳颖. 生成式人工智能创作使用作品的合理使用调适[J]. 科技与法律(中英文), 2024(3): 77-87.
ZHOU W K, FEI Y Y. Fair use adjustment of the use of works by generative artificial intelligence creation[J]. Science technology and law (Chinese-English version), 2024(3): 77-87.

[11] 张惠彬, 肖启贤. 人工智能时代文本与数据挖掘的版权豁免规则建构[J]. 科技与法律(中英文), 2021(6): 74-84.
ZHANG H B, XIAO Q X. The construction of copyright exemption rules for text and data mining in the era of artificial intelligence[J]. Science technology and law (Chinese-English version), 2021(6): 74-84.

[12] 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.
ZHENG F, XIA C B. The copyright dilemma and institutional response of generative artificial intelligence - Take ChatGPT and ERNIE bot as examples[J]. Science technology and law (Chinese-English version), 2023(5): 86-96.

[13] 林秀芹. 人工智能时代著作权合理使用制度的重塑[J]. 法学研究, 2021, 43(6): 170-185.
LIN X Q. Reshaping the fair use system in copyright law in the AI era[J]. Chinese journal of law, 2021, 43(6): 170-185.

[14] 林华. 大数据的法律保护[J]. 电子知识产权, 2014(8): 80-85.
LIN H. Legal protection of big data[J]. Electronics intellectual property, 2014(8): 80-85.

[15] 高阳. 衍生数据作为新型知识产权客体的学理证成[J]. 社会科学, 2022(2): 106-115.
GAO Y. Theoretical justification of derivative data as a new type of intellectual property object[J]. Journal of social sciences, 2022(2): 106-115.

[16] 冯晓青. 数据财产化及其法律规制的理论阐释与构建[J]. 政法论丛, 2021(4): 81-97.
FENG X Q. Theoretical interpretation and construction of data propertyization and its legal regulation[J]. Journal of political science and law, 2021(4): 81-97.

[17] 梅夏英. 企业数据权益原论: 从财产到控制[J]. 中外法学, 2021, 33(5): 1188-1207.
MEI X Y. On the interests on enterprise data: From property to control[J]. Peking university law journal, 2021, 33(5): 1188-1207.

[18] 崔国斌. 大数据有限排他权的基础理论[J]. 法学研究, 2019, 41(5): 3-24.
CUI G B. Towards a theory of limited exclusive right to big data[J]. Chinese journal of law, 2019, 41(5): 3-24.

[19] 冯晓青. 知识产权视野下商业数据保护研究[J]. 比较法研究, 2022(5): 31-45.
FENG X Q. Commercial data protection from the perspective of intelectual property rights[J]. Journal of comparative law, 2022(5): 31-45.

[20] 朱长宝. 论在线浏览、欣赏目的临时复制的法律保护[J]. 电子知识产权, 2016(10): 79-87.
ZHU C B. Study on legal protection of temporary reproduction of online browsing and appreciating[J]. Electronics intellectual property, 2016(10): 79-87.

[21] 张金平. 人工智能作品合理使用困境及其解决[J]. 环球法律评论, 2019, 41(3): 120-132.
ZHANG J P. Fair use of artificial intelligence: Dilemma and solutions[J]. Global law review, 2019, 41(3): 120-132.

DOI：10.13998/j.cnki.issn1002-1248.24-0475

[22] MURRAY M D. Generative AI art: Copyright infringement and fair use[J]. SMU science and technology law review, 2023, 26 (2): 259.

[23] 叶兆驰 . 人工智能生成物的侵权及解决路径[J]. 中南民族大学学报(人文社会科学版), 2024, 44(5): 156-163, 223.
YE Z C. Infringement of AI-generated works and resolution pathways[J]. Journal of south-central Minzu University (humanities and social sciences), 2024, 44(5): 156-163, 223.

[24] 潘香军 . 论机器学习训练集的著作权风险化解机制[C]//《上海法学研究》集刊2023年第6卷——2023年世界人工智能大会青年论坛论文集 . 香港: 2023年世界人工智能大会青年论坛论, 2023: 12.

[25] 邵红红 . 生成式人工智能版权侵权治理研究[J]. 出版发行研究, 2023(6): 29-38.
SHAO H H. Research on copyright infringement governance of generative artificial intelligence[J]. Publishing research, 2023(6): 29-38.

[26] BUITEN M, DE STREEL A, PEITZ M. The law and economics of AI liability[J]. Computer law & security review, 2023, 48: 105794.

[27] 刘小璇, 张虎 . 论人工智能的侵权责任[J]. 南京社会科学, 2018 (9): 105-110, 149.
LIU X X, ZHANG H. On the tort liability of artificial intelligence[J]. Nanjing journal of social sciences, 2018(9): 105-110, 149.

[28] LIOR A. AI strict liability vis-à-vis AI monopolization[J]. Science and technology law review, 2021, 22(1): 90-126.

[29] 朱阁 . “AI文生图”的法律属性与权利归属研究[J]. 知识产权, 2024, 34(1): 24-35.
ZHU G. A study on the legal attributes and ownership of "AI text-to-image"[J]. Intellectual property, 2024, 34(1): 24-35.

Copyright Data Dilemma of Building High-Quality Data System for AI: Present Situation, Coping Strategies, and Implementation Path

ZHANG Hecan[1], YI Chengqi[1*], GUO Peng[2], HUANG Qianqian[1,3], JIN Xiaokun[4]

(1. Department of Big Data Development, State Information Center, Beijing 100045; 2. Centre for Strategic Studies, Greater Bay Area Big Data Research Institute, Shenzhen 518048; 3. School of Information Resource Management, Renmin University of China, Beijing 100872; 4. Institutes of Science and Development, Chinese Academy of Sciences, Beijing 100190)

Abstract: [Purpose/Significance] Improving the policy and governance systems to promote the development of strategic industries such as artificial intelligence was explicitly proposed in the resolution of the Third Plenary Session of the 20th Central Committee of the Communist Party of China. In recent years, the conflict between AI companies' desire for copyrighted data and the copyright holders' protection of copyrighted data has become increasingly apparent. There have been a number of lawsuits and disputes around the world regarding copyright infringement caused by artificial intelligence. The dilemma of copyright protection of AI training data has become a difficulty and bottleneck that urgently needs to be resolved in the development of high-quality data system for AI. [Method/Process] Based on the academic research and industrial practice on the copyright protection of AI data, this study systematically summarizes six representative approaches to address the copyright dilemma of AI training data, and provides a comparative analysis of the advantages, disadvantages, and applicability of these approaches. The six representative approaches are: signing a license agreement by both parties, initiating special plans or forming alliances, introducing a copyright notice mechanism,

introducing a copyright risk guarantee mechanism, replacing with synthetic data, and applying copyright detection tools to large language models. For the copyright dilemma of AI training data, there is no optimal solution that can both encourage the supply of AI copyright training data and protect the copyright of data. [Results/Conclusions] In order to provide helpful references for increasing the supply of AI copyright data, formulating relevant policies, and promoting related work, this study has proposed a concept of general implementation path to build a high-quality data system for AI to solve the copyright dilemma of AI training data, based on the comparative analysis of the above six representative approaches and combined with China's four unique advantages. These include: 1) Integrating existing platforms to build a national-level integrated service platform for copyright data for AI, with state-owned enterprises (SOEs) under the direct administration of the central government taking the lead in establishing a national copyright data alliance and connecting copyright data to the platform. 2) To collaborate with local pilots of data intellectual property rights, explore and promote comprehensive reform pilot programs of copyright data adapted to the development of AI, and continuously strengthen the cooperation efforts and willingness between AI enterprises and copyright holders. 3) The focus should be on principled or critical issues, establishing and improving legislation related to copyright data for AI and promoting industry self-regulation.

Keywords: artificial intelligence; data system for AI; copyright protection; copyright data; data elements