Dispute Settlement 争议解决, 2025, 11(11), 49-55
Published Online November 2025 in Hans. https://www.hanspub.org/journal/ds
https://doi.org/10.12677/ds.2025.1111343

Hans汉斯

文本与数据挖掘的合理使用问题研究

邱敏敏

湖北大学法学院，湖北 武汉

收稿日期：2025年9月24日；录用日期：2025年10月27日；发布日期：2025年11月4日

摘 要

本文围绕文本与数据挖掘在人工智能研发过程中的著作权侵权风险展开研究，系统分析了TDM行为在复制权、改编权与信息网络传播权等方面可能构成的侵权类型。进一步探讨了将TDM纳入著作权法“合理使用”制度的理论与现实困境，包括与“作者中心主义”的冲突、利益平衡机制的破坏以及现有合理使用条款的适用局限性。在此基础上，结合欧盟、英国、美国的立法经验，提出在我国构建TDM合理使用制度的可行路径，包括引入“转换性使用”与“非表达性使用”概念、修改兜底条款为开放式一般例外条款等，以促进人工智能产业健康发展与国家创新战略的实施。

关键词

文本与数据挖掘，合理使用，著作权法，侵权风险

Research on the Fair Use of Text and Data Mining

Minmin Qiu

School of Law, Hubei University, Wuhan Hubei

Received: September 24, 2025; accepted: October 27, 2025; published: November 4, 2025

Abstract

This paper examines the copyright infringement risks associated with text and data mining in the development of artificial intelligence. It systematically analyzes potential infringements related to reproduction, adaptation, and communication to the public rights. Furthermore, it explores the theoretical and practical challenges of incorporating into the “fair use” (or exceptions and limitations) framework under copyright law, including conflicts with doctrines, disruptions to the balance of interests, and the limitations of existing fair use provisions. Drawing on legislative experiences from

文章引用: 邱敏敏. 文本与数据挖掘的合理使用问题研究[J]. 争议解决, 2025, 11(11): 49-55.
DOI: 10.12677/ds.2025.1111343

the EU, the UK, and the U.S., the study proposes pathways for establishing a TDM fair use system in China. Suggestions include introducing concepts such as "transformative use" and "non-expressive use", as well as modifying the current "catch-all" clause into an open general exception clause to support the healthy growth of the AI industry and align with national innovation strategies.

Keywords

Text and Data Mining, Fair Use, Infringement Risk, Legislative Comparison

1. 文本与数据挖掘的侵权风险

文本与数据挖掘的侵权风险核心在于，未经授权对受版权保护的大规模材料进行“复制、分析和输出”的过程中，可能触犯法律。

1.1. 人工智能研发训练阶段

自进入新世纪以来，高新技术不断涌现，人工智能凭借其无可比拟的应用潜力成为科技新星，被视为第四次工业革命的重要推动力。人工智能是通过计算机科学技术模拟人类智能的学科，广泛应用于互联网、云计算及各类数字平台。它显著解放了人类的脑力劳动，推动各行业迭代升级，大幅提升生产力。人工智能技术的深入应用正深刻改变生活方式，典型体现包括媒体机构采用 AI 生成新闻、歌手利用 AI 辅助创作、汽车借助 AI 实现无人驾驶等。在此背景下，科技企业积极布局人工智能领域，推出多种 AI 应用。例如美国 OpenAI 实验室开发的 ChatGPT，具备语言理解与文本生成能力，可在多场景中与人交互，完成邮件、脚本等文本撰写任务。

1.2. 数据获取与利用贯穿人工智能研发始终

人工智能是需要运营商通过大量的“培训”来实现的，运营商输入庞大的数据库，不断提高人工智能的模仿能力，甚至实现创新能力的突破。具体而言人工智能的运行大体可以分为模型训练和内容输出两个阶段，模型训练是人工智能产生的基础，在模型训练的基础上人工智能才能产生新的文本。模型训练包括数据收集、数据预处理以及构建数据集等环节，而数据挖掘属于人工智能模型训练必不可少的过程。数据挖掘是指在已知的数据集合中发现各种模型、概要和导出值，从大规模文本、数据中发现规律、趋势进而形成模型，实现自我提升的过程。可见人工智能创作呈现“数据输入－学习规则－创作产出”的过程。人工智能的进化与发展取决于三个要素：大数据、算法、算力[1]。文本与数据的输入是人工智能训练和学习的基础，也是人工智能发展的前提。只有通过海量的数据训练，人工智能才能找出规律，总结出模型和规则，可以说海量的文本与数据构成了人工智能的发展之基。

2. 文本与数据挖掘的可责性分析

人工智能的输出端与输入端的著作权问题引起了学界的广泛关注和讨论，从产出来，人工智能生成内容是否具有可版权性，能否受到著作权法的保护问题。从输入来看，人工智能训练所使用的数据是否构成侵犯。

2.1. “春风图案”与“奥特曼维权案”

2023 年 2 月份，李先生利用 Stable diffusion 人工智能大模型，通过输入提示词、设置相关参数生成了一张人物图片，该图片被命名为“春风送来了温暖”发布在网络平台。刘女士以该图片作为文章配图在个人账户发布，李先生认为其利用人工智能创作产生的图片构成作品，刘女士未经许可擅自使用的行为侵犯了自己的信息网络传播权，将其起诉到北京互联网法院。奥特曼是家喻户晓的动漫形象，原告获得奥特曼著作权人的独占授权，被告经营一家网站，提供具有 AI 对话及 AI 生成绘画功能的服务[1]。原告发现，可以要求网站生成奥特曼相关的图片，且生成的图片与原告奥特曼形象构成实质性相似。被告在未授权的情况下，擅自利用原告享有权利的作品训练其模型并生成实质性相似的图片，侵犯了原告对作品的复制权和改编权。

通过两个案例，可以发现人工智能渗透到了人们的日常生活，输入指令来生成文本、图片、视频已经司空见惯。在人们利用人工智能的同时，潜藏着巨大的侵权风险。人工智能相关的问题逐渐从学界的讨论到司法上的判决，在国外有纽约时报诉 OpenAI 案，我国的“春风图案”和“奥特曼案”等。

2.2. 文本与数据挖掘的侵权范围

人工智能利用的数据来源广泛，既包括公共领域的数据也包括个人领域的数据。公共领域的数据具有公开性和公众性，所有人均可任意使用，人工智能数据挖掘行为的相关利用也不会产生法律问题，而对个人领域数据的不当使用行为就可能导致侵权的发生。个人领域的数据根据性质的不同，可以分为人身性数据和财产性数据。人身性数据落脚于人格权，受个人信息法的保护。财产性数据主要指上传的数字作品，受著作权法的保护。本文主要讨论人工智能数据挖掘行为对著作权法上受保护作品的利用行为。与传统对作品的使用方式相比，人工智能的数据挖掘行为因数据的虚拟，而在作品的使用时表现明显的区别，体现在的规模性和隐蔽性。首先传统人工在创作时，利用的作品数量是有限的，在表达的时候也能体现出自己的独创性，即使原文引用也会标明作品的出处以及创作人的姓名[2]。人工智能的数据挖掘行为是大规模、大批量对作品的使用，人工智能通过数据训练的产物，具有含著作权作品的影子，这无疑会消解被使用作品潜在的价值，侵犯著作权人的权利。其次人工智能数据挖掘行为是在网络中按照设定的程序进行的，行为的主体并不明显。而且数据的抓取是在平台后进行的，难以对数据使用行为进行监督。

3. 合理使用制度对文本与数据挖掘的适用困境

人工智能的数据挖掘行为是对网络上的数字信息包括数字作品的大规模的使用行为，在未经著作权人的同意情况下，可能侵犯作者的复制权、改编权、信息网络传播权等相关权利，使得人工智能公司陷入到众多诉讼纠纷之中。为了我国人工智能健康的发展，能否通过著作权法上的制度来使人工智能数据挖掘行为摆脱所面对的侵权风险？

3.1. 合理使用制度适用于文本与数据挖掘的理论困境

随着人工智能逐渐成为重要创作主体，传统作者中心主义所主张的“作品源于作者”“作品是作者精神化身”等原则受到根本挑战。AI 生成内容是否构成作品，引发广泛争议。可以确定的是，在人工智能时代，作品创作不仅依赖于人类作者的能力，更受 AI 技术发展水平的影响，作者中心主义及其相关权利保护机制势必面临重大冲击。

合理使用制度是指在一定条件下对著作权人的专有权利进行限制的平衡机制，目的是平衡各方的利

[1]中国裁判文书网(2023)京 0491 民初 11279 号。

益，在保护著作权人权利的同时，也要兼顾公共政策的考量、保护言论自由、促进公共事业的发展等。对著作权人的限制有着严格的规定，《尼泊尔公约》《与贸易有关的知识产权协定》规定了三步检验标准，要求合理使用的范围是有限的，不能损害到作品的市场价值，且不能严重损害到权利人的合法利益，我国著作权法通过情况列举的方式对合理使用制度的适用严格限制了范围[3]。随着人工智能时代到来，数字作品因高度的流动性更容易被复制和使用，著作权人会陷入到维权困境，首先是难以发现侵权事实，其次是维权成本高且难以找到具体责任人。若将文本与数据挖掘纳入合理使用制度，那很有可能破坏公众与作者之间的利益平衡，即过度保护公众利益而严重损害到作者利益。

3.2. 合理使用制度适用于文本与数据挖掘的实践困境

著作权法旨在平衡著作权人专有权与公共利益，通过设置豁免制度防止权利滥用，主要包括法定许可和合理使用。法定许可允许不经权利人同意使用作品但须支付报酬，其适用范围由法律严格限定，如报刊转载、录制录音制品等，目前并不涵盖人工智能数据挖掘行为。即便未来纳入，由于 AI 训练使用作品规模巨大，适用法定许可仍将导致高昂成本，故难以适用。

合理使用则允许在特定条件下不经许可且无偿使用作品，但须注明来源且不得影响作品正常使用或不合理损害权利人合法权益。《著作权法》第二十四条列举了多种情形，如个人学习、适当引用、科学研究等。人工智能数据挖掘行为是否可纳入现有合理使用类型，仍需结合其具体使用目的、方式和影响进行个案判断，当前制度下仍存在适用困境。

《著作权法》中合理使用条款第一种情形：“为个人学习、研究或者欣赏，使用他人已经发表的作品”。可以看出，构成合理使用要求主体的目的是学习、研究或者欣赏等非商业性目的，人工智能的数据挖掘行为是为了训练人工智能，最终成果将上市面对公众，难以符合合理使用条款的目的要求。其次合理使用条款要求主体为个人使用，“个人使用”中的“个人”并不能当然延及个人的科研团队。人工智能开发作为一项系统性工程，显然单个个人无法承担这类科研的费用，也无法单独完成这样的科研任务。合理使用条款第二种情况为：“为介绍、评论某一作品或者说明某一问题，在作品中适当引用他人已经发表的作品”。此款规定构成合理使用的行为需满足介绍、评论作品或说明问题的目的，同时是“适当”的引用。数据挖掘行为的目的是为了训练人工智能来创作新的作品，而非说明某一问题。训练人工智能的挖掘行为需要海量的作品，且是对作品完整的使用，而非仅借鉴作品中的某一片段，人工智能创作使用数据不符合“适当引用”条款所要求的“适当性”要件[4]。在具备必要性的前提下，使用作品的数量、方式、范围还必须控制在一定的限度之内，因此人工智能数据挖掘行为不符合此条款对适当引用的要求。合理使用条款第六条规定：“为学校课堂教学或者科学研究，翻译、改编、汇编、播放或者少量复制已经发表的作品，供教学或者科研人员使用，但不得出版发行”。此条款规定了在科学研究情况下的对作品的合理使用，创作人工智能无疑属于科学研究，人工智能数据挖掘行为与此款规定最为贴切。但此款同时要求成果不得出版发行，可见利用作品的行为不能具有商业性的目的，因而该情形下的科研机构及科研活动应只适用于国家设立的教育、科研公共事业单位，利用作品创作人工智能无疑具备营业性。同时要求为科研人员使用，此处应当看作为个人使用，在人工智能公司的科研人员对作品的使用行为，应当看作为职务要求，为法人行为而非个人使用。因此人工智能数据挖掘行为难以归类合理使用条款中的科学研究情况。

4. 文本与数据挖掘纳入合理使用制度的正当性分析

近年来人工智能的发展得到世界范围内广泛国家的关注和重视，先后发布国家级人工智能战略政策，试图通过相关政策设计，对人工智能发展进行战略部署，从而占据有利地位，振兴国家的经济。

4.1. 人工智能产业发展前景与国家发展战略的考量

在新兴科技革命的背景下，各国通过加强技术合作和产业交流，积极发展人工智能的发展。如今人工智能产业已经初具规模，在我国人工智能得到迅速的发展和普及，通过与传统产业通过相互结合的方式，在一些领域之中已投入产业化运营，如无人驾驶技术，形成了新型的人工智能产业领域。同时，人工智能产业市场的总体规模逐年快速上升，对各国的经济发展做出了卓越的贡献。普华永道 2017 年的一份分析报告预测，截止 2030 年人工智能将为全球经济带来 15.7 万亿美元增长，对中国而言则将带来 7.0 万亿美元的经济增长，国内生产总值亦将增长 26.1%。2018 年国内人工智能产业市场总体规模达到 415.5 亿美元，较上一年度增幅达 75%。IDC 预测 2020 年全球人工智能市场规模为 1565 亿美元，较上一年度增幅达 12.3%；中国信通院数据研究中心测算 2020 年国内人工智能产业规模为 466.3 亿美元。从以上数据中可以看出，人工智能的产业与技术发展对我国经济做出了巨大贡献。近几年我国经济受到了重大冲击，为摆脱这一困局，促进产业升级和大力发展新兴产业日趋重要。为此我国已在此次新技术革命的进程中率先布局，占据了不少先发优势。无论是为了激励产业加大投入扩大发展，还是为了促进技术创新保障经济发展动能，都有必要为其扫清人工智能产业发展中所可能面临的制度障碍。进而维持我国技术与产业的稳步发展，巩固我国经济地位与能力。

4.2. 技术发展与著作权人之间的利益平衡考量

人工智能对著作权人市场利益的影响具有双重性：一方面，某些使用行为可能导致潜在市场收益的减损；另一方面，人工智能技术亦能带来积极效应，且不同类型的数据挖掘项目对权利人的影响存在差异。多数文本挖掘项目在使用作品后并不向公众提供完整或部分复制件，因而未直接损害作品欣赏市场的利益[5]。非生成式人工智能(如自动驾驶系统)的训练者既不公开训练数据，也不生成新的表达内容，故其使用行为通常不影响著作权人的市场收益。

尽管人工智能的开发可能对部分权利人的利益造成一定冲击，但其显著提升了社会整体生产效率与创作能力。生成式人工智能可辅助创作者降低生产成本、提高产出效率与质量，甚至帮助其获取更广泛的市场收益。例如，自然语言处理与机器学习技术能够分析海量素材、提取模式与趋势，为创作者提供灵感和风格参考，或直接生成诗歌、旋律与设计草图等初级创作内容。由此可见，人工智能的应用有助于大幅降低创作成本、减少人力投入、优化资源利用，并缩短创作周期，使创作者能够更灵活地适应市场需求、提高响应速度与产出水平，最终推动文化产业向更高效、多元的方向发展。

5. 文本与数据挖掘纳入合理使用的方案设计

英国和欧盟则对人工智能文本与数据挖掘合理使用认定增加了限制条件，如上表中的主体、客体、目的以及使用方式等方面的限制。对人工智能文本与数据挖掘的相关立法，各国需结合本土法律特色以及司法环境，因地制宜地选择合适的法律制度。

5.1. 文本与数据挖掘合理使用的立法模式选择

在全球人工智能治理的浪潮中，英国和欧盟对文本与数据挖掘(TDM)的“合理使用”例外并未采取“一刀切”的开放态度，而是为其套上了严谨的法律枷锁。它们通过立法，在参与主体、适用客体、合法目的以及合规使用方式等多个维度，设置了明确的限制条件例如，欧盟的《数字单一市场版权指令》为科学研究机构提供了更宽松的 TDM 例外，却允许权利人对商业主体的挖掘行为进行保留；同时，其对数据来源的合法性要求及对挖掘成果传播的限制，都体现了其旨在权利人与技术开发者之间寻求一种精密的平衡。这种立法模式，深刻地反映了欧盟在数字战略上兼顾创新与保护其深厚的版权产业利益的独特考量[6]。

5.1.1. 欧盟：公益目的的合理使用模式

2016 年，欧盟发布《数字化单一市场版权指令》，在版权制度上进行了一次重大变革，目的是为了应对信息技术的快速发展给版权制度带来的巨大挑战。在第三条规定科研机构和文化遗产机构为科学研究目的进行文本和数据挖掘，对其合法获取的作品或其他内容进行复制与提取的行为，属于权利的例外。第四条规定以文本和数据挖掘为目的，对合法获取的作品或者其他内容进行复制与提取的行为，属于权利的例外，但有前提条件，为权利人没有以适当的方式明确保留对上述作品或其他内容的使用。

通过分析法条，第三条数据挖掘行为的例外具有主体、目的、对象以及使用方式的要求。主体须为科研机构或文化遗产机构，而非所有的主体；使用目的为科学研究的非商业性目的，对象为合法途径获得的作品，包括纸质和电子形式的作品；使用方式为复制和提取行为，也就是说例外情况仅包含数据挖掘行为的基础阶段，不包含后续的改编和传播行为。可以看出此条对数据挖掘例外情况的适用有着严格的要求。

第四条为数据挖掘例外的一般情况，对主体并未有严格的限制，也未要求非商业性使用的目的，但规定了使用的前提条件，权利人未对作品进行明确保留使用，也就是权利人未明确禁止他人对作品进行数据挖掘性的使用。此条款进一步放宽条件，有助于促进人工智能产业的发展。但此条款在实践中也会存在些问题，权利人为了自己的利益，在发布作品的时候可能会注明禁止他人数据挖掘性的使用，当情况蔚然成风时，此条款也就流于表面，未达到立法的目的。即使欧盟对数据挖掘行为相关立法仍存在一些不足之处，但其应对数字技术带来的挑战，大胆进行法律修改的革新之举，对我国著作权法法律制度具有重要借鉴意义。

5.1.2. 英国：非商业性的合理使用模式

英国作为最早制定《版权法》的国家之一，是合理使用制度的开创国，也是通过立法方式确定文本与数据挖掘行为合法性的欧洲国家。为满足现实生活的需要，英国在 2014 年对《版权法》进行了新增修订，新增了文本与数据挖掘例外规则条款，当中允许为了非商业性研究的文本和数据挖掘目的，利用计算机分析技术对已经合法获得访问的任何版权材料进行复制。

通过法条分析，可以看出英国以立法的形式为文本与数据挖掘行为赋予了合法性，以防止版权成为阻碍人工智能技术发展的阻力。同时值得注意的是，英国同样为相关立法设置了许多限制条件。首先英国著作权法对行为“主体”设定限制，要求其为合法获取作品的主体，即行为人本身应当具备合法访问相关版权材料的资格。“合法获取”的方式通常包括订购数据、购买数据库等方式。其次英国对“使用目的”作出了限制，规定只有基于“计算机分析”和“非商业性使用”目的的文本与数据挖掘属合理使用范围，如科研人员进行的技术研究。同时此排除了不以计算机处理、分析数据为目的的行为及具有盈利性质的商业性使用。ChatGPT 作为营利性组织，其对数据挖掘、使用行为难以被定性为“非商业性使用”。英国著作权法在“使用行为”上进行了限制，人工智能数据训练的使用过程涉及著作权法上的多个行为，包括复制、改编与传播。而英国《版权法》第 29A 条只针对文本与数据挖掘的复制行为提供了合法性支持，对其他行为则未设置侵权豁免，因此相关行为仍存在侵权风险。

5.2. 引入“转换性使用”概念

“转换性使用”由莱瓦尔法官在《论合理使用标准》中首次提出，强调合理使用应具有生产性，须以不同于原作品的方式或目的使用作品，而非简单重包装或重发布。该理论主张通过增添新美学内容、新视角或新理念，使原作品在使用过程中获得新价值、功能或性质，从而转变其原有用途。在美国司法实践中，转换性使用成为合理使用判断的核心因素，尤其侧重于“使用的目的和性质”。在 Kelly v. Arriba Soft Corp 案中，法院认定被告搜索引擎提供缩略图的行为具有高度工具性与转换性，未替代原作品审美

功能，不损害其市场，因而构成合理使用。我国著作权法采封闭立法模式，仅明文列举十二类合理使用情形，缺乏如美国“四要素”般的开放弹性，尤其未引入转换性使用标准，导致难以适应技术发展带来的新型使用方式，面临类型化不足与制度僵化的问题。文本与数据挖掘不属于现有的合理使用情形，合理使用制度难以适用，在司法实践中，由于法条缺乏弹性，使得法官难以作出公平一致的判决，因此在我国引入“转换性使用”制度具有正当性。

5.3. 改造现行合理使用制度法律规范

针对人工智能数据挖掘行为的侵权问题，最简单便捷的方式为在《著作权法》的合理使用条款中，直接增设人工智能数据挖掘为合理使用的情形，这样直接豁免了作品使用人可能的侵权责任。当然，为维护使用者和著作人之间的利益平衡，有必要在规定数据挖掘行为为合理使用的同时，规定相应的前提条件或者限制范围。如欧盟在第四条规定在著作人未保留权利时，对作品的数据挖掘行为才构成合理使用。在第三条未设置前提条件，但对主体范围和使用目的有着限制。日本则对数据挖掘行为利用作品程度进行限制，要求为轻微程度的使用。我国在增设数据挖掘为合理使用情形时，可以规定人工智能开发者在进行模型训练时，利用他人享有著作权的作品为合理使用，但不得损害权利人的相关权益。在生成的内容中，若包含他人作品的，应当注明来源。

在《数字单一市场版权指令》的制定过程中，有观点主张欧盟应借鉴美国合理使用制度，引入开放式一般例外条款，以增强法律在数字与跨境环境下的适应性。美国版权法第 107 条规定的合理使用“四要素”标准，赋予法官较大的自由裁量权，能够灵活应对文本与数据挖掘等新兴技术带来的挑战，但也因依赖司法裁量而存在结果不确定性[7]。我国虽采封闭式立法模式，但司法实践中已体现出借鉴开放标准的趋势。最高人民法院曾在司法政策中指出，在促进技术创新和商业发展确有必要的特殊情形下，可综合考虑使用行为的目的、作品性质、使用比例及市场影响等因素，认定合理使用。这一思路为人工智能文本与数据挖掘等新型案件在司法实践中适用合理使用提供了可能，可将其纳入兜底条款中的“特殊情形”予以灵活处理。

参考文献

[1] 宋华健. 论生成式人工智能的法律风险与治理路径[J]. 北京理工大学学报(社会科学版), 2024, 26(3): 134-143.
[2] 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.
[3] 任庭汉. 创作型人工智能应用下作者“数据控制权”的提出[J]. 网络安全与数据治理, 2023, 42(9): 50-58.
[4] 吴汉东. 著作权合理使用制度研究[M]. 北京: 中国人民大学出版社, 2020: 235-236, 312, 312, 290.
[5] 万勇. 人工智能时代著作权法合理使用制度的困境与出路[J]. 社会科学辑刊, 2021(5): 93-102.
[6] 焦和平. 人工智能创作中数据获取与利用的著作权风险及化解路径[J]. 当代法学, 2022, 36(4): 128-140.
[7] 刘禹. 机器利用数据行为构成著作权合理使用的经济分析[J]. 知识产权, 2024(3): 107-126.