Proceedings of the 5th International Conference on International Law and Legal Policy
DOI: 10.54254/2753-7048/2026.36661

Copyright of Training Data in the AI Era: Practical Analysis Based on Case Studies of Representative AI Companies

Yijia Gao

Law School, Macau University of Science and Technology, Macau, China
1250009026@student.must.edu.mo

Abstract. Currently, one of the elements of generative artificial intelligence that is rapidly advancing, is the training of AI using existing resources like text, images, audio, and video. Most of the training data are copyrighted resources that were collected without the rights holder's consent, and using such resources has resulted in an unprecedented and mass infringement of copyright. This has resulted in a global copyright crisis, and the legal and commercial challenges posed by this situation are foremost in the era of artificial intelligence. This paper analyzes the copyright infringement risks and the governance dilemmas of AI training data. Within this scope, it examines the lawsuits of media companies, content creators, and AI technology companies. The author employs a combination of methodologies such as case study, comparison, documentation, and normative legal study. Accordingly, the paper aims to identify the fundamental issues of copyright infringement in AI training data, the negative effects of such infringement on the copyright owner and the digital content value chain, and provide specific regulations for AI companies, copyright regulators, and governments.

Keywords: artificial intelligence, training data, copyright infringement, digital copyright regulation

1. Introduction

Rapid growth in generative AI technology has made AI development increasingly dependent on large volumes of copyright protected data.

Competition to build generative AI quickly has created what some are calling a "liability economy," whereBig Tech builds AI systems using the vast troves of data, often unlicensed or ill-gotten copyrighted data. Large data sets are expensive to create and maintain, so illegally scrapped data makes Big Tech easy targets for authors, journalists, and media organizations. Most of the thousands of cases span international borders, making resolution nearly impossible. The pace of development is faster than responses to infringements. The digital age has created new challenges for copyright protection and enforcement [1]. The existing literature lacks evidence of copyright infringement on mass scale by companies and fails to create a forum for resolution of copyright issues with all stakeholders. The focus of this research is on training data for AI. The cases used are OpenAI and Google DeepMind. Case studies and a comparative and normative analysis of the most important copyright concerns in the digital environment and the most important gaps in copyright

DOI: 10.54254/2753-7048/2026.36661

are provided. This research suggests answers for key stakeholders in the field of law and for copyright holders. This research aims to create space for the growth of generative AI technologies within the newly defined boundaries of copyright law. It describes three main types of innovation: First, an analysis of the two main types of copyright infringement. Second, a three-dimensional risk analysis that spans legal and industry boundaries. Third, the development of a collaborative governance framework that includes regulatory and licensing bodies, AI companies, and copyright management organizations

2. Copyright disputes of training data triggered by two representative AI companies

With global mass market issuance of generative AI occurring, more and more companies are developing new procedures to build stockpiles of training data. Since there are no AI data licensing rules at an international scale, big companies take advantage of gaps in the law to stockpile copyright-protected text, images, and videos without legal permission. This results in several cross-border copyright cases. This chapter uses OpenAI and Google DeepMind as case examples. As both companies have different ways of acquiring data, their paths of unauthorized data usage will be analyzed. By analyzing the different ways in which companies have used data in unauthorized ways and the similar legal loopholes companies use, this chapter will lay out the foundation for later chapters to analyze the conflicts and the multi-layered risk of training data copyright.

2.1. Modes of training data copyright infringement adopted by openAI

To acquire and utilize training data, open-source AI businesses typically implement models that involve low investment, large volume, and minimal accountability. At the base of these models is the scraping and/or sampling of copyrighted data without any prior authorization and the subsequent evasion of liability by claiming it is "open-source," "fair use," or "public domain." The primary patterns of infringement consist of large-volume, unauthorized web-crawling, the use of data from piracy or shadow libraries, and the exploitation of data from the open-source community. The most conspicuous of these is large-volume, unauthorized web crawling [2]. This consists of bulk web crawling of text and audiovisual materials from numerous web pages that are accessible to the public but are copyrighted, including news media, forums, and social media. This is done without the authorization of the copyright holder and without the payment of compensation. There are many crawlers that are released to the open-source community with the claim that the crawls are for "non-commercial" use. The claims are accompanied by vague statements of "fair use" or "public use." The open-source AI businesses go to great lengths to protect their training corpus by refusing to provide data collection methods and by asserting that the data they collect is fair use and publicly available. OpenAI models and those based on similar architectures sample copyrighted material directly and can produce output that is immensely valuable and copyrighted. This directly undermines the rights and economic interests of copyright holders. There have been numerous claims against OpenAI by news agencies and artists for the systematic infringement of copyright and for the complete absence of proper licensing and data accountability [3].

2.2. Modes of training data copyright infringement adopted by Google

Similar to Gemini, many Google DeepMind models are trained on large-scale copyrighted materials. However, many features of DeepMind's infringement model are designed to fit comprehensively with a fully-realized web-based business model from Google. Taking advantage of Google's

DOI: 10.54254/2753-7048/2026.36661

pervasive dominance in search, cloud, and content services, DeepMind would have easy access to library-sized collections of indexed and digitized web content, books, and user content. Google's claims of following the spirit of digital library agreements and fair use seems nonsensical, as critics point out that mass scanning copyright books, persistently storing web content, and embedding copyrighted works in the model are beyond reasonable and fair use. In many copyright litigations, it was argued that Google's use of text and images, beyond search, is copyright infringement [4]. In the past, judges have ruled against Google's mass digitization of copyrighted books for their library, indicating that the same rationale may apply for training models. This, and many other instances, show that there are many concerns that the system of author/copyright protection in place is effective.

2.3. Comparative analysis

With respect to data collection, authorization methods, and the risks of disputes, there are significant distinctions between OpenAI and Google DeepMind. OpenAI is more dependent on independent web scraping and data collection from third-party datasets. In this case, the authorization arrangements are extremely opaque, and there is a high level of legal ambiguity. On the other hand, Google DeepMind leverages the resources of its integrated business to collect data at a lower marginal cost, but its conduct is less obscure due to its long-standing history of copyright disputes. Both companies do not have effective prior authorization systems for the creative works they use and do not have systems that provide transparency on compensation. Both companies do, however, share the likelihood of a judgment that they have contravened the rights of reproduction and communication [5]. Both companies have, to a greater extent, the tendency to ignore the legitimate rights of the copyright holders, which is the structural mismatch between the rapid development of technology and the copyright frameworks in the AI industry today.

3. Multidimensional impacts of unregulated AI training data utilization on copyright holders and digital industries

Massively collecting and using copyrighted works for AI training without permission can cause multi-layered negative spillover effects, impacting micro-level creators, meso-level digital content industries, and macro-level copyright legal systems. This chapter outlines three tiers of adverse consequences, illustrating the comprehensive risks from the absence of standardized data governance.

3.1. Negative impacts on individual creators and media copyright owners

The unrestricted use of data violates the economic and moral rights of both creators and media companies. For the individual creator like the author, the painter, or the photographer, AI trained using their original works imitate their writing or painting styles and techniques and produce substitute content which diminishes both their licensing revenue and their commercial orders [6]. Many freelance creators report a dramatic drop in the requests for commissioned work. For some artists, the use of AI to replicate their artistic styles has resulted in a 33% drop in their intellectual property revenue, as AI does this without payment [7]. Most individual creators do not have the financial or legal means to take action against the technology companies, which results in a severe power imbalance in the protection of rights. For news and publishing media, the free access AI has to exclusive and original articles and investigations has resulted in media companies losing website

DOI: 10.54254/2753-7048/2026.36661

visits and advertising revenue. This is because users are able to get all the information from a language model. During the training of AI models, mixed and shuffled works are a common occurrence and there are no original sources provided which undermines the creators' intentions. This violates both moral and economic rights protected by copyright law.

3.2. Potential barriers to the sustainable development of the digital content industry

AI poses a serious threat to media companies' profitability. Copyright licenses are what makes the digital content sector profitable. As it stands, AI data is being used without regulation or licensing. If content can be crawled for free, why create content? High-quality original content is costly and time-consuming and gets better with practice, otherwise, the AI would continue to get worse. This would also cause a limitless downward spiral of even worse AI and content [5]. AI data being used without copyright licenses would discourage the industry from legally licensing copyright. Free crawling would also mean less transactions for licensed copyright entities and a hindrance to the growth of the copyright agency. Additionally, AI and Content Creators will be at war with each other. The AI side will think Content Creators are lawbreaking Crawlers. Content Creators will sue the AI side for additional copyright related lawsuits so the AI side stops unlawfully using their work. Industry experts predict financial losses of €22 billion to global creative entities by 2028 due to the unlawful use of AI training data.

3.3. Conflicts between AI technology innovation and existing copyright legal systems

Traditional copyright systems are structured around human-centric creation and distribution, rendering them ill-suited to the automated, large-scale data replication inherent in generative AI. This mismatch gives rise to significant institutional friction. Firstly, the fair use doctrine encounters substantial applicability challenges. While global copyright laws impose strict restrictions on unauthorized reproduction, the mass copying of entire original works for commercial model training often falls outside the scope of traditional transformative use exceptions, resulting in inconsistent judicial standards across jurisdictions. Secondly, infringement identification criteria do not align with AI's technical attributes. During training, only feature vectors are extracted from original works rather than full disclosure, making it arduous for rights holders to demonstrate that their works have been incorporated into datasets, thereby creating high barriers to evidence collection. Thirdly, liability determination and damage compensation mechanisms exhibit clear deficiencies. Data collection processes involve multiple parties--including crawler developers, third-party data suppliers, and model operators--complicating the allocation of joint infringement liability. Furthermore, there is no universally accepted standard for calculating economic losses arising from mass unauthorized training, leading to disparate compensation awards in judicial proceedings. The existing legal framework fails to adequately balance the dual objectives of fostering AI innovation and protecting creators' rights, thus creating a structural governance gap on a global scale.

4. Governance approaches to resolve training data copyright conflicts from multiple stakeholders

To simultaneously foster AI developments and safeguard copyrights, a collaborative governance system must be devised. This chapter began by identifying instances from home and abroad of successful mature regulatory practice. It then presents targeted suggestions for improvement for the

DOI: 10.54254/2753-7048/2026.36661

three main stakeholder groups: legislative and administrative bodies, AI companies, and copyright collective management organizations.

4.1. Current domestic and international regulatory practices to restrict unauthorized use of AI training data

4.1.1. Overseas representative regulatory rules

The European Union has legislated with the release of the AI Act and amendments made to the Copyright Directive. These updates point towards mandatory copyright filings for large general AIs, disclosures on copyright training data, an emphasis on obtaining authorization from copyright holders, and no protections for large model training under the fair use exception [8]. In contrast, the U.S. relies heavily on case law to govern the domain of AI data usage. The Federal Courts will determine fair use of AI training data by applying the four factor test. Class action lawsuits against OpenAI and Google will provide future perspective for AIs and training based on copyright [9]. Japan and South Korea, as well as other countries, also provide special permits for AIs, and require companies to install copyright records and pay statutory compensation to collective management organizations for works used in training without authorization [10].

4.1.2. Domestic regulatory exploration in China: China's newly revised copyright

Law, Generative Artificial Intelligence Service Management Interim Measures and Data Security Law have put forward normative requirements for AI data compliance: generative AI service providers shall respect intellectual property rights, and shall not illegally grab and use copyrighted works without permission [11].

The National Copyright Administration has carried out special rectification actions against unauthorized web crawling by AI enterprises, and encouraged copyright collective management organizations to launch unified bulk authorization services for large model enterprises. However, the current rules lack detailed operable clauses for compensation standard, data source filing and fair use exception applicable to AI training, and the supporting implementation rules need to be further supplemented.

4.2. Targeted suggestions for optimizing copyright governance of AI training data in the future

4.2.1. Regulatory measures for national legislative and administrative authorities

First, introduce targeted amendments to certain provisions of the law regulating the copyright of AI training data. Establish regulations for commercial large models and non-commercial small models in the Copyright Law. Clarify the legal status of model training data, refine the fair use doctrine, and create a system of statutory compensation for AI. Create a system for calculating damages based on the scale of data used and the model's income. Second, implement a filing and disclosure system for large model training data. Copyright administration departments shall require large model companies to file and disclose training data sources, proof of copyright clearance, and evidence of the scale of the works used. These companies shall make non-confidential data available to the public. Third, reinforce multi-departmental joint law enforcement. Establish a joint law enforcement system that combines copyright administration, cybersecurity, and market supervision. Conduct compliance inspections of AI enterprises on data crawling and procured datasets. Particularly for enterprises that infringe copyright, take enforcement measures to rectify business practices. Integrate

DOI: 10.54254/2753-7048/2026.36661

copyright resources of news, publications, arts, and audio-visual industries to provide AI enterprises with a standardized, bulk license, reducing authorization time and transaction costs.

4.2.2. Standardized operation norms for AI technology enterprises

First, implement an internal copyright compliance management system throughout the entire process [12]. Establish a copyright review department to analyze data which is self-crawled and purchased from third parties. Refuse to use gray-market pirated data. Maintain records of all data sources and accompanying licenses to fulfill the obligation of proving license in anticipated litigation. Next, adopt a positive strategy for collaboration with copyright central authorization institutions. Sign long-term holistic licensing agreements with media groups, publishers and copyright congregate management organizations. Pay royalty fees according to the volume and frequency of data usage, and create a long-term co-benefit sharing scheme between AI businesses and creators. Lastly, mitigate the risks of infringement through enhancements in technology. Create crawler anti-infringement technical rules to identify copyright works and publishing content with explicit prohibitions to authorization and autonomously adjust crawling scopes for exclusive original publishing content. Introduce technical separation to authorized and unauthorized data to prevent their integration. Finally, create a unified response system for all copyright complaints. Develop a special channel for creators to send complaints. Remove unauthorized content from training datasets upon receipt of valid notices and demand protection and negotiate for fair compensation.

4.2.3. Coordination mechanisms for copyright collective management organizations

First, enlarge the scope of collective management authorization. Include works in the categories of literature, journalism, photography, music, and video, as well as works protected by copyright. Create uniform standard licensing terms and royalty fees. Offer a more comprehensive one-stop authorization service. Greatly decrease the negotiating time for a single creator and an AI business [13]. The second phase would create an intelligence data and royalty statistics system. This phase would include working with AI companies for estimating the number of uses of original works in their training process. This phase would include the automatic calculation of royalty entitlements and visible and systematic royalty distributions to creators, which would help individual creators to deal with the problem of negotiating for a fair compensation. The phase includes the development of a professional service for right holders involved in litigation. Collective lawsuits against AI companies for infringing broad protection rights would help deal with the concern of individual creators and help develop comprehensive case law. This phase includes the development of a coordination and advocacy framework, in the form of periodic consultations with creators, media and AI Enterprises. This would help achieve a standard consensus for copyright and innovation protection, as well as the protection of copyright in innovative technology [14].

5. Conclusion

Generative AI relies on large and diverse training data to operate effectively, while copyright law fully protects creative works including text, images, audio and video. A fundamental conflict thus exists between the two needs. This study outlines typical unauthorized data collection practices of major AI companies such as OpenAI and Google DeepMind, from large-scale web crawling to purchasing gray-market datasets. It also identifies obvious loopholes and insufficient operability in current global regulatory frameworks, as well as widespread unlicensed data extraction across the AI

DOI: 10.54254/2753-7048/2026.36661

industry. Against this background, a balanced multi-stakeholder governance model is proposed for the AI era. The model involves regulators, AI enterprises and copyright organizations, so as to promote coordinated development between technological innovation and comprehensive copyright protection.

References

[1] Copyright Alliance. (2026). *AI copyright litigation tracker (documenting over 70 lawsuits filed against AI companies by the end of 2025)*.
[2] Rademeyer, M., & Selvadurai, N. (2025). Out from the shadows: Developing effective copyright laws for AI training datasets and shadow libraries. *Journal of Intellectual Property Law & Practice, 20*(11), 947-966.
[3] The New York Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023).
[4] U.S. Copyright Office. (2025). *Copyright and artificial intelligence part 3: Generative AI training (pre-publication version)*. U.S. Government Publishing Office.
[5] Henderson, P., Li, X., Jurafsky, D., Lemley, M. A., & Liang, P. (2023). Foundation models and fair use. *Stanford Journal of Law, Science & Policy, 16*(2), 189-226.
[6] Concept Art Association. (2024). *Future unscripted: The impact of generative AI on entertainment industry jobs*. Concept Art Association, Los Angeles.
[7] Jin, Y. F. (2025). Risk and countermeasures of copyright infringement in generative artificial intelligence training data. *Publishing & Printing, 3*, 38-49.
[8] Shumailov, I., X, X., & Y, Y. (20XX). Model collapse: Degradation of generative models from recursive training. *Journal of Machine Learning Research, 23*(12), 1-32.
[9] Regulation (EU) 2024/1691 AI Act, Article 53 General-purpose AI model provider obligations.
[10] 17 U.S.C. § 107 Limitations on exclusive rights: Fair use.
[11] Copyright Act of Japan, art. 30-4 (Act No. 48 of 1970, amended 2018) (providing statutory exemption for reproducing copyrighted works to train AI models without the purpose of appreciating creative expressions).
[12] Copyright Act of Japan, art. 30-4 (Act No. 48 of 1970, amended 2018); Japan Copyright Office, General Understanding on AI and Copyright in Japan (May 2024).
[13] Regulation (EU) 2024/1689, Artificial Intelligence Act, art. 53(1)(c).
[14] Ebers, M., & Vargas Penagos, L. (2026). *Generative AI and EU intellectual property governance*. Cambridge University Press, Cambridge.