Proceedings of the 4th International Conference on Global Politics and Socio-Humanities
DOI: 10.54254/2753-7048/2026.33835

Copyright Infringement Risk and Governance of Generative AI Training Data

Yi Zhang

College of Law and Politics, Hebei GEO University, Shijiazhuang, China
zy18733061028@gmail.com

Abstract. About the copyright infringement problem of generative artificial intelligence, the core is how to balance the demand of technology development for massive data and the protection of the legitimate rights and interests of copyright owners. Although various mainstream programs proposed by the current academic circles have their own theoretical contributions, they all face insurmountable institutional obstacles in practice. This paper proposes a collaborative governance path based on the core concept of "incentive-sharing". Advocates are abandoning the single system dependence. Building a multiple collaborative system of "market mechanism leading, collective management supporting, and rational use of supplement". In terms of specific regulations, it is suggested to implement a relatively loose authorization mechanism for the input end of the model. Establish a benefit compensation mechanism of "giving priority to voluntary licensing, supplemented by extended management" through collective management organizations. At the output end, strict standards for identifying substantially similar infringement are adhered to prevent the generated content from replacing the original work in the market. This path aims to achieve a dynamic balance between encouraging creation and promoting sharing. Providing institutional support for the copyright law ecology in the era of artificial intelligence.

Keywords: Generative Artificial Intelligence, Training Data, Copyright Infringement, Motivation and Sharing

1. Introduction

In February 2024, the Guangzhou Internet court entered into force a judgment on the infringement of others' copyright by generative AI services. This is also the first effective judgment on the infringement of others' copyright by generative AI services in the world. This decision brings the copyright issue of generative AI training data from theoretical deduction to practical controversy. Almost at the same time, cases such as the New York Times v. OpenAI and France imposing hundreds of millions of euros on Google emerged one after another. Indicating that the copyright dispute of artificial intelligence has evolved from a theoretical discussion to a realistic challenge that needs an urgent response.

The core of the problem is that generative AI needs a large amount of work as training materials. It is not feasible to obtain the authorization of copyright owners one by one. If free use is allowed, it will seriously damage the interests of creators and eventually lead to the exhaustion of high-quality

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2026.33835

works for training. In the face of this dilemma, the academic community has put forward diversified solutions, which have their own advantages and disadvantages. Based on the systematic review of existing research, this paper proposes a collaborative governance path with the core concept of "incentive-sharing". It provides theoretical support and practical guidance for the improvement of the copyright system in the era of artificial intelligence.

2. Infringement risk in the whole process of training data

In the stage of work collection, the main copyright risk in the data utilization of such diverse resources is the difference in the ownership, authorization, and licensing of data. The copyright resources that have exceeded the protection period are in the public domain and can be used by anyone without barriers. There are still some disputes about the use and payment of ownerless works in practice. For copyright works still in the protection period, there is a certain risk of copyright infringement in the acquisition and collection stage of copyright content [1].

In the work input link, it is necessary to convert the captured work into a machine-readable standard data format for cleaning, marking, etc. To form a new data set corresponding to the demand, and realize the targeted input of data content. This link will involve unauthorized infringement of the translation right, adaptation right, and compilation right in the copyright of the work [2].

In the content output stage, generative AI may output content similar to the style and expression of other people's works, grab the interests of the copyright owner. Causing the interests of the copyright market to be unbalanced. The copyright owner is unable to obtain due economic benefits from the infringing works, which dampens the enthusiasm of the creator. The intellectual achievements of creators cannot be properly protected, which will undermine the balance of copyright ecology and have a negative impact on the development of the entire cultural industry [3].

3. Review of existing research: theoretical contributions and practical limitations of multiple programs

3.1. Fair use theory: value orientation and boundary of innovation priority

The theory of fair use is one of the most influential propositions at present. Supporters believe that the use of works by AI training belongs to "non-expressive use". Whose purpose is to extract the facts, styles, and patterns in the works, rather than reproduce the original expression of the works. So it should not constitute infringement. The relatively loose identification standard at the data input end will help to provide a legal source path for the development of the AI industry.

However, the theoretical defects of fair use theory are also significant. A single regulation on the free use of works for training commercial AI systems will damage the economic interests of copyright owners. AI-generated works not only have the appearance of human works, but also their content has approached or even surpassed human works in quality. If they are allowed to flow into the market, they will inevitably occupy the market share of copyright owners [4]. This will make the copyright owner unable to get returns from the emerging use of works, dampen his creative motivation to rely on works for profit, and is not conducive to encouraging creation.

DOI: 10.54254/2753-7048/2026.33835

3.2. Collective management system: institutional advantages and practical obstacles of centralized authorization

The copyright collective management system is regarded as an effective mechanism to solve the problem of licensing massive works because of its "one-to-many" centralized licensing function. The establishment of an artificial intelligence training data works management organization can not only efficiently help the users of work to obtain due benefits, but also reduce the legal risks of artificial intelligence enterprises and reduce litigation disputes. Through the centralized management of rights to promote the use of rights, it reduces the search cost of users, and also promotes the use of works. While solving the problem of the dispersion of rights, it does not shake the exclusivity of copyright [5].

However, China's copyright collective management system itself still faces problems. The existing coordination mechanism among the five collective management organizations is not perfect, and it is difficult to realize the one-stop authorization of cross-category works. In addition, the pricing mode, distribution mechanism, publicity channels of non-member works, and other supporting systems that only use the training data usage fees of the system still need to be further explored.

3.3. Compensation system: technological innovation and right compensation system innovation

The copyright compensation system advocates that the state should give proper compensation to copyright owners through taxation and other ways to balance the relationship between industrial development and rights protection. The scheme tries to go beyond the system adjustment within the copyright law. To realize the secondary distribution of interests with the help of fiscal and taxation tools. Supporters believe that the compensation system can effectively solve the double dilemma of authorization efficiency and rights protection cost. For AI enterprises, it eliminates the burden of negotiation one by one. For obligees, they can share the dividends of technology development without actively monitoring infringement.

However, critics point out that although the system helps to improve the economic situation of the author group, it cannot provide economic compensation for the individual author. Even though, in theory, the tax paid by AI enterprises can be directly paid to the author in a fair and reasonable way through system design. Such a system scheme still has the same problems of information asymmetry and implementation cost as the legal license theory [6].

4. Collaborative governance: institutional adjustment of generative AI copyright protection

Throughout many research programs, it is not difficult to see that a single application of a certain program is difficult to meet the current needs. Therefore, based on the evaluation of existing research, this paper proposes a collaborative governance path with the core concept of "incentive sharing".

4.1. Concept innovation: the necessity of collaborative governance

The so-called "incentive" means that the system design should fully respect the incentive function of the copyright law to ensure that creators can obtain reasonable returns from the emerging use of works. The so-called "sharing" refers to reducing the transaction cost of the use of works through institutional innovation, and promoting the wide dissemination and efficient use of knowledge resources in the era of artificial intelligence. "Incentive" and "sharing" are not antagonistic, but

DOI: 10.54254/2753-7048/2026.33835

complementary. Only when the creators get continuous incentives can high-quality works be produced continuously, and the training data of artificial intelligence will not be exhausted. Only when works are used efficiently can copyright owners share the dividends of technological development and form a virtuous cycle of "creation - utilization - return - re creation".

4.2. Hierarchical regulation: from "single system" to "collaborative governance"

The core proposition of the collaborative governance path is that the copyright protection of artificial intelligence should not rely on a single system. It should build a multiple collaborative system of "market mechanism leading, collective management supporting, and rational use supplement". Different systems play their respective advantages in different scenarios, forming a governance network with complementary functions.

The specific implementation is first reflected in the differential regulation between the input and output. The relevant guiding opinions of the Supreme People's Court and judicial practice have preliminarily formed a consensus. Focusing on promoting the acquisition and use of data in the model pre-training stage, pay more attention to the protection of rights in the generation and dissemination stage. This system design idea of "Leniency in and strictness out" has sufficient legal basis.

At the input end, the Hangzhou Internet Court pointed out in the Altman case. "In view of the purpose of using other people's works in the data training phase of generative AI, in principle, it should be used to learn and analyze the thoughts and feelings, language features, characteristic styles, and other contents expressed in the previous works. Extracting the corresponding rules, structures, patterns, trends to facilitate the subsequent transformative creation of new works" [7]. Such use "is not for the purpose of reproducing the original expression of the work", which can be considered as reasonable use if the conditions are met. As the core defense of copyright infringement, it brings the data training behavior of generative AI into the scope. This not only conforms to the institutional presupposition and value objectives of rational use, but also encourages data utilization and digital technology innovation [8].

However, the loose input terminal does not mean free use. For commercial AI systems, a reasonable interest compensation mechanism should be established. Based on this, this paper proposes to build an authorization system with collective management organizations as the hub and "voluntary licensing as the main, supplemented by extended management". For the works of the members of the collective management organization, the collective management organization and the artificial intelligence enterprise shall negotiate to determine the authorization agreement. For the works of non-members, the extended collective management system can include them in the scope of management, and give the right holder the right to "opt out" at any time. The composite standard of "basic fee+floating fee" can be adopted for the collection of royalties. The basic fee is determined according to the scale of training data. The floating fee is linked to the commercial income of the artificial intelligence system. It is allocated by the collective management organization according to the actual use frequency, contribution, and other factors of the work. At the same time, it is necessary to increase the transparency of collective management and eliminate the influence of various human factors, which is conducive to the protection of copyright owners' right to know. Irregular audits can also be adopted, and third-party institutions can be hired to audit copyright collective management organizations. So as to effectively prevent possible problems in the process of collective management [9]. In some training scenarios involving a large amount of fragmented data, such as the training involving social media data, it is difficult to identify the specific copyright owner due to the wide dispersion of data sources. At this time, through this system, data trainers can

DOI: 10.54254/2753-7048/2026.33835

use this data for training under the premise of legality, and copyright owners can also obtain corresponding compensation from the compensation fund [10]. This mechanism not only protects the right to remuneration of the obligee but also avoids the high cost of individual negotiation.

At the output end, it should adhere to the strict standards of infringement identification. At present, it should go beyond the existing criteria for judging substantive similarity. In the specific comparison, it can combine the expression elements of the same author or similar works as an important consideration angle for the overall observation of substantive similarity judgment in the judicial trial. So as to break through the traditional dilemma of "style thought" category and bring deep imitation into the category of substantive similarity [1].

5. Conclusion

The copyright governance of generative AI is essentially an institutional practice that seeks a dynamic balance between technological innovation and rights protection, private interests and public welfare. The numerous schemes proposed in the existing studies have their own limitations, which enlighten people to transcend the limitations of a single system with the thinking of collaborative governance.

The collaborative governance path proposed in this paper takes "incentive-sharing" as the core idea. Constructing an authorization compensation mechanism with hierarchical regulation at the input and output ends and collective management as the hub. The legitimacy of this path lies in responding to the reasonable demand of the artificial intelligence industry for massive data. It also ensures that creators can share the dividends of technological development and maintain the core function of copyright law to stimulate innovation through the coordination of multiple systems.

The value of the collaborative governance path is that it is based on the basic principles of copyright law and explores the possible space for institutional innovation with a pragmatic and rational attitude. Only by seeking a balance between incentives and sharing, and keeping a balance between protection and opening up, can build a copyright legal ecology that promotes technological innovation and protects the rights and interests of creators, and meet the knowledge prosperity in the era of man-machine coexistence.

References

[1] Dai, J.L. and He, R.N. (2025) Large Model Training Data Copyright Infringement Risk Regulation. Digital Library Forum, 21(9), 74-81.
[2] Wei, L.L. (2025) The Applicable Choice of Copyright Restriction System in Generative AI Data Training. Treatise on Politics and Law, (5), 159-172.
[3] Huang, M.S. (2025) Copyright Risk and Governance Path of AI Large Model Training Data. Journal of Hubei University (Philosophy and Social Sciences), 52(5), 185-193.
[4] Zhang, X.C. (2026) The Legal Logic and Implementation Path of the Rule of Fair Use in Generative AI Data Training. Journal of Chongqing University of Science and Technology (Social Sciences Edition), (1), 55-65.
[5] Yao, Y. (2025) Is It an Elegant Crime to Steal Books by Machine?--Identification and Risk Mitigation of Copyright Infringement of AI Generated Content. Jinan Journal (Philosophy and Social Sciences Edition), 47(8), 71-89.
[6] Cao, Q.Z. (2025) Copyright Protection of "Free and Open Source" Generative AI Training Data. Journal of Renmin University of China, 39(6), 88-99.
[7] A Cultural Development Co., Ltd. in Shanghai v. A Smart Technology Co., Ltd. in Hangzhou for Copyright Infringement and Unfair Competition Dispute, Civil Judgment of First Instance of Hangzhou Internet Court (2024) Zhe 0192 Minchu No. 1587.
[8] Shao, W.Y. and Huang, Z.Y. (2026) Analysis on the Rational Use of Copyright of the Training Data of Generative Artificial Intelligence Model. Journal of Wuhan University of Technology (Social Sciences), 39(1), 103-110.

DOI: 10.54254/2753-7048/2026.33835

[9] Hua, J. (2025) Research on Copyright Collective Management License of Generative Artificial Intelligence Model Training. Lanzhou Academic Journal, 1-12.
[10] Nie, H.T. and Hou, J. (2026) Generative AI Data Training Copyright Infringement Risk and Regulation Path. Journal of Xihua University (Philosophy and Social Sciences), 45(1), 11-22.