WEP Warwick Evans Publishing

International Journal of Social Sciences and Public Administration

ISSN: 3005-9836 (Print), ISSN: 3005-9585 (Online) | Volume 8, Number 2, Year 2025
DOI: https://doi.org/10.62051/ijsspa.v8n2.13
Journal homepage: https://ijsspa.org

An Exploration of Training Data Compliance from a Copyright Law Perspective

Lingli Yang

School of Law, Beijing Normal University, Beijing, China

ABSTRACT

AI technology, including generative artificial intelligence such as data training, has become a key force in developing new quality productivity. Currently, the copyright controversy in large model training has attracted much attention both at home and abroad. Starting from the legality of data source at the data input end and the compliance of data use at the output end, the copyright risk of the big model training process can be comprehensively deconstructed. Diversified dispute resolution mechanisms such as the copyright exception system, the safe harbor principle and the collective management system have faced many challenges in practice, especially the traditional authorization and licensing model has shown obvious limitations in the face of the huge volume of data and the large number of right subjects. We analyze the trend of extra-territorial regulation and explore the feasible paths for solving data training copyright disputes in China, i.e., guided by the principle of balancing interests, promoting the construction of a consultation mechanism between the right subjects and AI enterprises; implementing a data labeling system; drawing on the open source licensing in the field of computers; and broadening the boundary of fair use.

KEYWORDS

Data Training; Copyright; Data Compliance; New Quality Productivity.

1. PROBLEM FORMULATION: COPYRIGHT CONTROVERSY IN LARGE MODEL TRAINING IS IN THE SPOTLIGHT

On March 5, 2024, the government work report emphasized "vigorously promoting the construction of modern industrial system, accelerating the development of new quality productivity". [1] As a typical representative of new productivity, AI big models have reshaped the logic of content creation and empowered people's work and life, and at the same time triggered a series of new challenges in the copyright system. Whether the training data is compliant or not is the starting point for analyzing all AI copyright issues. On the one hand, clarifying the copyright dispute of data training process is conducive to ending the uncertainty of copyright infringement in the development of large models. On the other hand, unauthorized data training copyright dispute cases are increasing. According to incomplete statistics, from November 2022 to October 2023, only the U.S. District Court for the Northern District of California has accepted 10 cases of copyright holders v. Open AI and other AIGC R&D enterprises for unauthorized use of copyrighted works to conduct model training. [2] In June 2023, there was also a dispute in the online teaching industry in the country for unauthorized use of third-party platform works data for large model training. resulting in strife. On the list of defendants, various Internet giants could be seen among them.

In the current technology field, although the differences between different models at the algorithmic level gradually tend to be small and show a trend of homogenization, the importance of training data is becoming more and more prominent, and it has become the core element that distinguishes and

determines the performance of large models. Lack of data means that the training of large models cannot be carried out effectively and continuously advanced. Taking Open AI's several generations of GPT models as an example, the training data parameters have increased from 5GB in GPT-1 to 100 trillion in GPT4, and the demand for data has soared exponentially. Enterprises rely only on their own data and public data resources is far from enough, from the Internet and other friends that "hitchhiking" has become an open secret in the industry, but also become a trigger for the frequent occurrence of copyright litigation.

Behind the frequent litigation is the contradiction between the big model's extreme thirst for data and the unsound data compliance mechanism, the traditional copyright law can't adapt to the way of data use in the AI era, which triggers new problems in the application and supervision of the law.

2. RESPONSIBILITY FOCUS: COPYRIGHT RISK ANALYSIS OF DATA TRAINING

2.1. Analysis of the Legality of Data Sources at the Input End of Data Training

Article 7 of the Interim Measures for the Administration of Generative Artificial Intelligence Services (hereinafter referred to as "the Measures") issued by the State Internet Information Office stipulates that: "Generative artificial intelligence service providers shall, in accordance with the law, carry out pre-training, optimization training, and other training data processing activities, and use legally-sourced data and basic models; they shall not infringe on the intellectual property rights enjoyed by others in accordance with the law. Where personal information is involved, the consent of the individual shall be obtained or in accordance with other circumstances stipulated by laws and administrative regulations; ....." The core objective of this provision is to clarify the data compliance obligations of generative AI service providers, which is also the core topic of the global regulation of the legitimacy of AI training data.

The legality of the data source is a necessary condition for the compliance of generative AI products. If the data source is not legitimate and compliant, the product will not be able to enter the market application stage. However, if excessive emphasis is placed on the review of data source legitimacy during the R&D stage, it may compress the scope of big data collection, slow down the speed of technology iteration, and thus restrict the output effectiveness of highly intelligent products. Therefore, Article 3 of the Measures also emphasizes a basic principle: to promote the organic combination of innovation and the rule of law while ensuring that development and safety are given equal importance, to support the innovative development of generative AI through the adoption of effective measures, and to implement an accommodating, prudent and categorized and graded regulatory attitude towards its services.

The issue of legitimacy of training data sources occurs in the data collection process. In order to expand the sample data set and improve the quality of sample data, the collector will collect data as comprehensively and extensively as possible. The legitimacy of the data source of generative AI mainly refers to whether the data collected by generative AI is obtained in a lawful and legitimate manner, whether it does not jeopardize the rights and interests of the data rights holder, whether it obtains the consent of the relevant intellectual property rights owner, and whether it obtains the consent of the subject of the personal information or other data rights holders when processing the personal information, and so on. This paper focuses on the analysis of data source legitimacy from the perspective of copyright law.

As a complex collection of data texts, the ownership of training data is complex and is mainly divided into three categories: public domain data, authorized licensed data and unauthorized data. Although public domain data does not involve copyright property rights, there is still a risk of violating the author's right of attribution, the right ofintegrity of the work and other moral rights. Authorized data

usually refers to content obtained through one-to-one licensing agreements with copyright holders or through legal authorization by copyright collective management agencies, but this part of the content is also usually difficult to achieve 100% accurate authorization. [3]

Unauthorized use of data is the most controversial and carries the potential risk of copyright infringement. Large model companies generally obtain sample data through four methods: "crawler" technology to capture data, deciphering technology to copy text, digitizing technology to reproduce information, and adding mandatory licensing clauses to user agreements. Among them, as a mainstream data collection technology, the legality of "crawler" technology needs to be comprehensively determined in combination with the terms of the specific crawler agreement, the collection objectives and the use scenarios, and commercial application scenarios usually face a higher compliance threshold.

The mere collection of data does not change the information and meaning of the data content, and is an act of reproduction in the de facto sense. After the Supreme People's Court defined scanning and digitizing as a method of reproduction through a judicial interpretation in 2000, Article 10 of the Copyright Law as amended in November 2020 added "digitizing" as a means of exploitation to the definition of reproduction right. As a result, the act of training data collection is included in the scope of the discussion of reproduction right.

2.2. Analysis of the Legality of Data Use at the Output End of Data Training

1) Direct use

Based on research, testing and other basic functions of the data training behavior does not involve the risk of copyright infringement, the direct use of such training data is not discussed.

The direct use in this paper refers to the behavior of using the collected data sets with copyright disputes directly to the data output without any technical Whether direct use of training data infringes on copyright needs to be judged from the source of the data and the purpose of its use. If the training data is from the public domain (a work whose copyright has expired) or if its author has permitted its use under some kind of open license (e.g., the Creative Commons license), then the use of the data may not constitute copyright infringement. Also important is the purpose for which the training data is being used. In some cases, academic research may be considered "fair use," especially if the purpose of the research does not involve commercial gain, affect the market value of the work, etc. However, this "fair use" may be considered to be "fair use. However, the boundaries of such "fair use" are blurred and vary depending on the legal system and the circumstances. Commercial use is usually more difficult to justify as fair use and therefore carries a higher risk.

2) Use after technical processing

In order to improve the quality of training data, companies usually clean and process the collected training data to remove duplicates, supplement data and correct erroneous data. Data cleansing is the process of repairing, correcting and organizing raw data to ensure data quality prior to data analysis or machine learning. Data cleansing is usually an iterative process that may require reviewing and improving the data several times until the analysis or modeling needs are met. However, this does not completely avoid the risk of copyright infringement for the following reasons.

For one, de-duplication is not the same as removing copyright infringing content during the cleaning process. [4] De-duplication refers to screening to eliminate duplicate content in the sample dataset to ensure that each data point is unique, rather than screening the training data for copyright-infringing content. This aspect of the process does not circumvent the risk of copyright infringement.

Second, data transcoding during the cleaning process may infringe on adaptation rights. [5] Training data transcoding is a broad concept that encompasses format conversion, encoding conversion, and content conversion in three ways, often with the goal of matching the data to a particular application,

model, or system. Format conversion changes the way data is stored and organized. For example, converting a CSV file to JSON format or converting an image from JPEG format to PNG format. Encoding conversions convert data from one character set encoding to another, such as from GBK encoding to UTF-8 encoding, in order to address compatibility or information sharing issues. Neither of the first two types of transcoding materially change the content and are not a concern. In the case of content conversion, the transcoded data may differ in some way from the original data. For example, images may lose part of the information after compression or cropping; text data may remove some noise or irrelevant content after preprocessing.

According to China's copyright law, the right of adaptation is an innovation in the form of the original work rather than a change in the substantive content. Whether the transcoding of training data will substantially change the output content depends on the purpose and method of transcoding. Thus, the content conversion in the transcoding technology may change the substantive information of the data content, thus infringing the right of adaptation.

3. FEASIBILITY ANALYSIS OF COPYRIGHT EXCEPTION SYSTEM

The copyright exception system includes the exceptions to ordinary exclusive rights such as fair use, compulsory license and statutory license, as well as the exceptions to specific exclusive rights such as the exceptions to the reproduction right, the exceptions to the translation right, the exceptions to the right of dissemination of information network, and the exceptions to the circumvention of technological measures. [6]

The process of processing the training data in depth involves multiple layers of conduction operations. Textual data is analyzed frequently, including multiple "copies" of the data. Such copies are usually made for the purpose of temporary storage and instantaneous reproduction of the data, and their period of existence is only measured in seconds, and these copies will automatically disappear when the computer is shut down, so they are called temporary copies. [7] Currently, in China's academic and legal consensus, temporary copying is regarded as a necessary technical operation, which itself does not have independently measurable economic value, and therefore should not be regarded as an infringement of the right to copy. Temporarily copied data exists only in the initial stage of data collection and is not researchable, while the data remaining after screening is the training dataset worth exploring.

Thus, the main focus of this paper is to explore the feasibility of authorization licensing system and fair use rules.

(1) Licensing System

Theoretically, the copyright licensing system can exempt the infringement liability. However, as far as data training is concerned, it is basically impossible to realize in practice. The reason is that the volume of training data is huge, and it is difficult to determine the subject of the rights, not to mention obtaining the authorization of the right holders one by one. In addition, China's "Copyright Law" for the "statutory license" provisions are more scattered, summarized mainly include "periodicals reproduced," "literary and artistic groups performances" and "sound recordings to produce phonograms". In addition, the provisions of China's Copyright Law on "statutory licensing" are scattered, summarized as four categories, including "reprinting in journals", "performance by literary and artistic groups", "production of phonograms by phonographic recordings", "production of radio and TV programs by radio and TV stations using the published works of others", which are very different from the model training behaviors, and it is difficult to match the application of the model. Under the background of AI big model, the traditional authorization and licensing model has failed. [8]

First, the right basis of authorization is unclear. The core of authorization is the creator's right to control the work. However, in the big model data training scenario, the definition of such control is

complicated by "model memory". Machine learning is an iterative and dynamic process, and the AI model has "learned" the features of the data during the training process and incorporated this information into the model's parameters. This means that the model "remembers" the data to a certain extent, and that information about a particular data point is dispersed and embedded in a number of parameters, even if the data no longer exists directly in the training set. Therefore, as the training model changes, the creator has no control over the work, and there is no license. Furthermore, the data involved in AI training is usually a collection of a large number of independent works. These works may come from many different creators, and each work may have different authorization status and conditions. From the external appearance, data training is similar to the re-creation of a natural person's study of a work, but in fact, it is not possible to specifically correspond to the established exclusive rights of copyright. The reason is that the use of data in the AI training process is not "use" in the traditional sense. Instead of directly consuming or redistributing the content, it analyzes the data and learns from it. The specificity of this use is in many cases beyond the scope of the original authorization, making it difficult to apply the exemption mechanism based on the existing authorization license.

Second, the feasibility of the authorization is questionable. As mentioned earlier, the training dataset may contain thousands or even millions of independent data points, and obtaining explicit licenses for each of them is almost impossible. [9] To optimize model performance, the development of AI models often requires rapid iteration, which means that the training dataset needs to be constantly adapted and updated. With prior authorization licensing, each dataset update may require a new round of licensing, which is very inefficient in practice. In addition, certain types of AI training require real-time access to dynamically updated data (e.g., streaming data), and it is even more difficult to achieve prior authorization in this scenario. Even if the problem of large-scale data authorization can be solved technically, legal and ethical issues are a major obstacle. Differences in copyright laws in different jurisdictions, privacy protection regulations, and restrictions on cross-border transmission of data may all impose limitations on the use of data. Even if data can be authorized for use, the challenge is how to ensure that its use meets ethical standards (e.g., does not exacerbate bias, does not infringe on individual privacy, etc.).

Third, authorization may lead to negative effects, potentially inviting "chilling effects", "data silos", "model bias" and other ills. Obtaining all the necessary authorizations can lead to significant increases in legal and administrative costs, especially for complex AI projects that require large and diverse datasets. This could make such high costs affordable only for well-resourced industry giants, limiting participation and innovation by smaller startups or academic institutions. Even when companies do their best to manage licenses, there is still a potential risk of infringement due to the complexity of data sources and rights holdings, which could directly lead to a "chilling effect" on the development of AI technology and industry. If each organization adopts a closed-door approach, using data only for internal purposes and setting thresholds for other researchers or developers, the problem of data silos may arise. This not only hinders data sharing and collaboration opportunities, but may also lead to duplication of efforts and waste of resources among different organizations. In addition, a restrictive licensing system may inhibit the potential for innovation and lead to a lack of diversity in training data, counteracting the problem of bias in AI systems. [10] For example, data from certain user groups may be less likely to be included due to cost or availability, leading to a decrease in the performance of the model when processing data from these groups.

Fourth, the royalty system is not yet practical. A royalty system, if applied to the training of large models, can promote the process of data compliance and ensure that creators can receive revenue and compensation from the reuse of their works. However, there are many obstacles at the practical level. First, it is difficult to assess pricing. Determining the exact amount of royalties for each work will be extremely challenging, as different data has different values for AI training and is difficult to quantify. The value of the training data depends not only on the work itself, but is also affected by factors such as the way it is used, the scope of use, the size of the dataset, and the expected commercial value. The

impact of different data on the training quality of an AI model also varies, with some potentially extremely critical and others almost negligible. Therefore, establishing a fair, transparent and efficient copyright pricing mechanism faces great difficulties. Second, it is difficult to track and manage. Each rights subject needs to negotiate royalties, and each transaction needs to be recorded, audited and enforced, which would be a large and complex administrative task. In order to achieve accurate royalty payments, highly complex technological solutions may need to be developed to automatically track data usage, calculate royalties, and execute payments. [11] The development and deployment of such technology is a huge challenge in itself. Models may need to be updated or retrained, and it is equally challenging to build a royalty payment mechanism that can adapt to such dynamic usage patterns and be effective over time. [12] Third, legal regulation is difficult. Different countries and regions have different legal regulations on copyright protection and royalty systems, which poses a challenge to establish a globally uniform royalty mechanism for training data.

(2) Fair Use Rule

The rule of fair use is embodied in Article 24 of China's Copyright Law, and the specific rules related to the training of AIGC models include "personal use", "proper citation", "use for study and research", etc. "and so on. [13] Personal use" is limited to non-commercial use, while the AIGC model is a commercial application for the public, which is inconsistent with it; "appropriate citation" presupposes that the work introduces, comments on or explains a certain issue, while the commercial application of the AIGC model can hardly be categorized under these circumstances; at the same time, "appropriate citation" can hardly be classified as a "commercial application"; and "study and research use" can hardly be classified as a "commercial application". Meanwhile, "scientific research" limits the use of the work to teaching or scientific research activities and emphasizes that only a "small amount" of reproduction can be made, which is inconsistent with the large-scale reproduction and use of the work by the AIGC model.

Although the Copyright Law as amended in 2021 added the "general elements" and "other circumstances stipulated by laws and administrative regulations" to the provision of "reasonable use", it is regarded as a half-assed provision. However, it is regarded as a semi-open provision and there are difficulties in its application. The rule of fair use is highly dependent on the circumstances of the case. This means that even in seemingly similar situations, courts may rule differently. The general elements and the underpinning provisions still need to be interpreted in terms of their applicability and scope through case-specific hearings, thus creating difficulties in prediction and judgment and increasing uncertainty in judicial practice.

Therefore, whether the AIGC model training can be included in the "fair use" exception, we still need to wait for the Copyright Law, the Copyright Enforcement Regulations and other relevant laws to be amended to make clear provisions.

4. PATH PROSPECT: ACCELERATING THE EXPLORATION OF THE EXEMPTION MECHANISM OF LARGE MODEL TRAINING LIABILITY

(1) Implementing a data labeling system

Data annotation is the manual or semi-automatic labeling of features or targets in raw data so that machine learning algorithms can learn from them. Data annotation usually requires manual participation because many features need to be intuitively felt by a natural person, such as object categories in images and speech recognition in audio. For features that are not amenable to direct manual annotation, such as sentiment analysis in text and attack detection in network traffic, semi-automated annotation methods are required. Although enterprises can use some tools and algorithms to assist in labeling data, due to the inflexibility of the technology, semi-automated labeling methods are far inferior to manually labeled datasets of higher quality. [14] For example, in natural language

processing, automatic annotation tools can be used to label lexical properties for text, but the annotation results still need to be corrected manually.

The Provisions on the Administration of Deep Synthesis for Internet Information Services emphasize strengthening the management and security of training data. Article 8 of the Interim Measures for the Administration of Generative Artificial Intelligence Services refines the specific requirements on data training: setting up clear, detailed and enforceable data labeling rules in line with these Measures; carrying out assessment of the quality of data labeling, and verifying the correctness of the labeled content through random sampling; and training for labeling personnel. The National Data Bureau said at the first national data work conference that it would carry out a pilot data labeling base. The top-level design and the system on the ground jointly promote the data labeling process. Data annotation gradually realizes autonomous recognition capability by adding recognition labels to images in advance and driving computers to learn the process through feature extraction. Therefore, the accuracy of data labeling determines the degree of intelligence of artificial intelligence to a certain extent, and the specific rules on data labeling made by the interim measures are conducive to improving the degree of intelligence of generative artificial intelligence, which in turn enhances the "accuracy" of training data.

Academic paper writing as an example, in order to avoid the risk of copyright infringement, to comply with academic standards, the author will inevitably have to cite the literature labeling, labeling is an effective means of exemption from copyright infringement liability. Although the form is different, the principle is the same. In the input side of the data training, after the collection, screening, cleaning of the data set by manual identification of copyright infringement, supplemented by technical tools and algorithms to mark the content of copyright protection, in order to avoid the risk of copyright infringement. However, due to the high degree of professionalism and the huge amount of sample data, the team of marking personnel needs to grow, and the requirements for their specialization are higher, which may bring greater employment costs to enterprises. But in the long run, it is a feasible strategy. There are three reasons: First, machine learning is based on the original database is constantly updated and developed, labeled data sets will become a new sample library of machine learning, so it basically avoids the duplication of labor of the labeling staff; Second, the training data samples are only a little more than that, with a huge volume of database without copyright exemption protection for enterprises is tantamount to embracing a gradually expanding time bomb. Compared with the uncertainty of infringement and the potential huge compensation, the cost of labor is not worth mentioning; Third, increase the positions of labeling personnel, which is beneficial to alleviate the pressure on the job market.

(2) Explore a new type of copyright exception system

The open authorization mechanism originates from the open source license in the field of computer software, which is a kind of open authorization statement of copyright to the world, through which the license specifies that the user is free to use, modify and share the content under certain conditions. This model effectively breaks through the limitations of the traditional one-to-one authorization model and improves the efficiency of resource use and accessibility. With the development of time, the open licensing mechanism has gradually expanded to various content and work areas such as documents, pictures, audio and video. During the evolution of the open licensing mechanism, the scope of application has covered documents, images, audio and video and other types of works. Under this framework, the Creative Commons license (CC license for short) builds a new paradigm for traditional copyright licensing of works, which recognizes copyright and respects the authors' right of authorship while allowing works to be used free of charge for personal use and commercial use, and allowing users to modify the works and redistribute them under certain conditions. The mechanism establishes the principle that the work can be freely exploited provided that the user complies with the conditions explicitly set out in the license, whereas in the event of a violation of these conditions, the open authorization is immediately terminated and the use of the work becomes subject again to traditional intellectual property protection. [15] It is important to note that the original

author or copyright holder is not responsible for the risks and consequences that arise from the use of the work.

In the field of artificial intelligence, open source licenses have become an important cornerstone of software technology development. AI software commonly adopts open source protocols such as the MIT license, the BSD license, and the Apache license, and these licenses support rapid iteration and widespread sharing of technology. To further advance the development of AI, the construction of training databases also needs to draw on this open source culture. By providing a free and flexible way to use copyrights, the open licensing mechanism not only promotes the development of the software and technology fields, but also provides a new way for the dissemination and innovation of cultural and scientific works. Today, with the rapid development of artificial intelligence technology, the value of this mechanism is becoming more and more prominent, and is expected to promote wider knowledge sharing and application innovation.

5. CONCLUSION

With the booming development of artificial intelligence technology, the issue of training data compliance has gradually become a focus that cannot be ignored. The future intelligent society needs a legal system that can protect the passion for creation and the fruits of labor without hindering scientific and technological innovation and social progress. Exploring this issue from the perspective of copyright law is not only a response to the challenge of adapting to the existing copyright legal framework, but also a forward-looking approach to the construction of the legal order of the future intelligent society.

REFERENCES

[1] Government Work Report, https://www.gov.cn/yaowen/liebiao/202403/content_6939153.htm

[2] Pamela Samuelson. "Generative AI meets copyright." Science (2023). 158 - 161.

[3] Tao Qian. On the Protection of Artificial Intelligence Generated Results by Copyright Law - Evidence of Data Processor's Right as Neighboring Right[J]. Law,2018,(04):3-15.

[4] Liu, Youhua,Wei, Yuanshan. Copyright infringement problem of machine learning and its solution[J]. Journal of East China University of Political Science and Law,2019,22(02):68-79.

[5] Jiao Heping. Copyright risks and path to resolution of data acquisition and utilization in artificial intelligence creation[J]. Contemporary Law,2022,36(04):128-140.

[6] Ma Zhiguo,Zhao Long. The Impact and Response of Text and Data Mining on the Copyright Exception System[J]. Journal of Northwest Normal University (Social Science Edition),2021,58(04):107-115.DOI:10.16783/j.cnki.nwnus.2021.04.012.

[7] Yu J,Sun Xuejing. Copyright law regulation of temporary reproduction in edge computing[J]. China Publishing,2022,(07):46-52.

[8] Wu Handong. Copyright Law on Works Generated by Artificial Intelligence[J]. Chinese and foreign law,2020,32(03):653-673.

[9] Zeng Tian. Research on Copyright Infringement Problems of Artificial Intelligence Creation[J]. Hebei Law,2019,37(10):176-189.DOI:10.16494/j.cnki.1002-3933.2019.10.012.

[10] Xiong Q. Copyright Recognition of Artificial Intelligence Generated Content[J]. Intellectual Property Rights,2017,(03):3-8.

[11] GUO Yan. Difficulties and Cracks in the Copyright System of Artificial Intelligence Creations[J]. China Publishing,2018,(12):67-69.

[12] LI Jing,ZHANG Jingchen. Implications of the Music Modernization Act for China's copyright system - copyright collective management organizations, music databases, royalty calculation[J]. Journal of Guizhou Normal University (Social Science Edition),2021,(01):139-150.DOI:10.16614/j.gznuj.skb.2021.01.016.

[13] Yao Ye. An Exploration of the Issue of Legislative Anomie in the Fair Use System in the Context of Digital Technology: A Concurrent Review of Article 24 of China's "Copyright Law" [J]. Technology and Publishing, 2021, (03): 140-145. DOI: 10.16510/j.cnki.kjycb.2021.03.007.

[14] Hu Mingyu, Xia Xue, Yang Chenxue, et al. Design and Implementation of a Semi-Supervised Image Annotation System Based on Deep Learning [J]. Journal of China Agricultural University, 2021, 26(05): 153-162.

[15] Zhu Hongtao, Niu Xiaohong. An Analysis of the Reasons for Differences in Authorization under Knowledge Sharing License Agreements: Based on the DOAJ Platform [J]. Modern Information, 2020, 40(05): 140-147.