HiLL PUBLISHING

Journal of Humanities, Arts and Social Science, 2025, 9(3), 573-576
https://www.hillpublisher.com/Journals/jhass/
ISSN Online: 2576-0548
ISSN Print: 2576-0556

Construction of Statutory Licensing System for Generative Artificial Intelligence

Yuxin Du

College of Marine Law and Humanities, Dalian Ocean University, Dalian 116023, Liaoning, China.

How to cite this paper: Yuxin Du. (2025) Construction of Statutory Licensing System for Generative Artificial Intelligence. *Journal of Humanities, Arts and Social Science*, 9(3), 573-576.
DOI: 10.26855/jhass.2025.03.024

Received: February 13, 2025
Accepted: March 11, 2025
Published: April 9, 2025

***Corresponding author:** Yuxin Du, College of Marine Law and Humanities, Dalian Ocean University, Dalian 116023, Liaoning, China.

Abstract

In the context of the rapid development of generative artificial intelligence (AIGC), the fair use system faces many challenges, especially in the process of data mining and machine learning, and its applicability is significantly limited. The fair use system is based on the "three-step" judgment standard proposed by the Berne Convention, which requires that the use of a work must be carried out under specific and non-universal circumstances, and must not harm the normal use of the work or infringe the legitimate rights and interests of the author without cause. However, generative artificial intelligence often fails to meet these conditions through the collection and processing of massive data. For example, data cleaning and sorting in the machine learning stage may affect the normal use of the work, while the behavior of data mining may also infringe the market value of the work. In addition, even under the framework of the "four-step judgment method" in the United States, fair use is limited to non-commercial research and development, which is obviously not conducive to the innovation and development of the increasingly commercialized generative artificial intelligence.

Keywords

Intellectual property rights; generative artificial intelligence; statutory license; fair use; machine reading

1. Research background

The fair use system in traditional copyright law is facing a breakdown when dealing with generative AI. According to the Copyright Law, fair use must be for non-profit purposes such as "personal study and research", and the usage proportion must be "small". However, the pre-training subjects of generative AI are mostly commercial companies, with a profit-making purpose, and they need to fully replicate works to improve model performance, far exceeding the limits of fair use. Scholars have proposed expanding the interpretation or adding new provisions to include AI training in fair use, but such solutions have been criticized as "going beyond the literal meaning of the law" or "ignoring the interests of copyright holders". In addition, fair use prioritizes fairness and is difficult to solve market failure problems, while the statutory licensing system prioritizes efficiency and reduces negotiation costs by paying a fixed fee, and is regarded as a more feasible alternative solution.

2. The main body of the generative AI statutory licensing system

The determination of the remuneration obligation entities under the statutory licensing system of generative artificial intelligence is a complex legal issue involving multiple links and entities. In the operation process of generative artificial intelligence, it can mainly be divided into three stages: machine reading, machine learning, and machine output. In practical applications, this process is jointly constituted by data suppliers, data training parties, and service

application layers. Data suppliers collect data sets and create them into general or specialized data sets. Data training parties mark and optimize the data sets, and service providers use their technical capabilities to meet customer needs and generate content. During this process, multiple entities have used the works. From the perspective of civil law, "each right corresponds to a corresponding obligation", so each entity seems to should bear the obligation of remuneration. However, this simple inference is not fully applicable. It must be analyzed based on specific circumstances.

Specifically, there are two main ways of text and data mining. The first way is that the training party of generative artificial intelligence directly acquires Internet information through crawler software and conducts data collection without going through any other entities. In other words, the training party is the creator of the data set. The second approach is that generative AI trainers purchase datasets for model training. For the first case, the data training party has control over the sources and usage of the data during its application, and thus should bear the corresponding remuneration obligations for the works involved in the data. In the second case, the data supplier and the data user are not the same subject, resulting in the separation of the "creator" and "user" of the data. This lays the foundation for the data set creator's obligation to pay: First, in this case, the creator of the data set collects and produces the data set and provides it to the trainer for use, while the trainer of generative AI purchases the data set for model training. Since statutory licenses typically involve users of a large number of works, and rights holders are dispersed and difficult to contact individually, paying each right holder directly by the user can result in high transaction costs. Therefore, the data set creator, acting as an intermediary, centrally handles the collection and distribution of remuneration, which can greatly reduce the cost of negotiation between the user and the right holder. By centralizing negotiations, data set creators ensure that they get their fair share, avoid being marginalized in the market, and ensure that data trainers have access to the datasets they need at a reasonable price. For this reason, when a data trainer buys a data set to train, the obligation to pay is the creator of the data set, not the data trainer. Secondly, from the perspective of behavioral content, the "basic model layer" and "professional model layer" have strong control over the source and application of data in the technical operation, so they should bear the corresponding payment obligation (Zhang Shulin & Wang Jianyu, 2024). The basic model layer generates general models through the learning of a large number of works, while the professional model layer makes domain-specific optimization on this basis. In contrast, the "service application layer" does not directly involve the training or processing of work, and its main role is to invoke the trained model to provide users with services such as text generation and image generation. Although the service application layer does not bear the obligation of payment, it still needs to bear a certain duty of care for the legitimacy of the copyright source of the large model it relies on to provide services and the legitimacy of the output content. Finally, at the normative level, the introduction of the "contact" requirement of generative AI means that as long as the generated work does not constitute plagiarism, subsequent authors do not need to confirm the existence of the previous work (Chu Meng, 2021). Therefore, if the data trainer obtains the data by means of crawling, it will become the subject of payment obligation because of direct contact with the work; However, for the data set acquired by the data trainer through purchase, because it has indirect contact with the work in essence, it does not need to become the subject of payment permitted by law, and the payment obligation should be borne by the creator of the data set.

To sum up, in the statutory licensing system of generative artificial intelligence, the subject of payment obligation should be determined according to the method of data acquisition, the role of the user, and its actual control over the work. When the data trainer obtains data through the crawler, it should undertake the obligation of payment; And when the data trainer buys the data set for training, the obligation to pay is borne by the creator of the data set. In addition, although the "service application layer" is not directly involved in the training process, it still bears some responsibility for the legality of the generated content. These analyses provide a theoretical basis for the improvement of the generative AI legal licensing system.

3. Object determination of the generative artificial intelligence statutory license system

3.1 Differences between general data sets and proprietary data sets

General data sets are characterized by diversity and comprehensiveness, usually covering a large number of public domain works or information that does not constitute works. The value of general data sets is reflected in the huge data volume and multidimensional data structure. Its content mainly includes: works that have entered the public domain and information that does not constitute works. As for public works, works in the public domain are not protected (Art Neill, 2024). As for the information that does not constitute a work, its main feature is that it is difficult to call it a "work" because of the triviality and one-sidedness of the data that constitute the data set, and it is closer to

the category of information in essence. In contrast, a specialized dataset is a data set built for a specific field or topic, such as the works of a specific author, the creation of a specific historical period, the performance of a specific artistic style, or the data of a specific industry. Such datasets are typically smaller in size than general-purpose datasets, but their value stems primarily from the professionalism of the data, the integrity of the work, and the high quality of the work. By their very nature, the materials that make up such datasets are often referred to as "works."

The significance of distinguishing between general data sets and proprietary data sets is whether the general data set should be the object of legal permission.

3.2 Build a "data exclusivity" model

With the rapid development of generative AI, how to train and solve copyright problems in a legal framework has become a key issue. Due to the highly diverse and fragmented information content covered by the general data set, it is difficult to fully reflect the idea of the work, so it cannot be completely called a "work", and there is no need to pay fees to the right holder or ask for consent. However, proprietary data sets contain a large number of copyrighted works and can fully reflect the author's thoughts. The construction process of these proprietary data sets will lead to the creation of a market for relevant right holders instead of a statutory licensing system in the process of their development and use, so as to balance the author's exclusive right and the user's right of contact.

In order to further promote the development of the field of generative AI, it is particularly important to build a reasonable business model. Through contractual cooperation, rights holders or third-party organizations can create proprietary content repositories for AI to use, both to ensure that copyright owners receive their due compensation and to avoid the risk of copyright infringement by research organizations during training. For example, Elsevier databases provide the content of journal and book chapters in XML format to support text and data mining activities, and users can access and use the content through APIs. The flexible market is already ahead of the law, and many companies are already working with rights holders to develop new business models that not only protect the interests of copyright holders but also avoid the copyright risks that research institutions face when using data. Google has proposed to the Authors Guild that it create a "Book rights registry" to manage the licensing of digital works in one place, with rights holders able to opt out if they do not agree (Wang Wenmin, 2022). Although this scheme has not been adopted, its rationality and feasibility provide a way to solve the copyright problem in the future.

From the purchase of data to the construction of an exclusive model, the copyright cooperation mode of generative AI is gradually taking shape. Data purchasing is an effective way to solve this problem, with dedicated data providers collecting, collating, and selling data according to demand, ensuring that companies use data legally and in compliance. Whether passive or proactive, providers of generative AI services have developed a licensing model for obtaining copyright licenses through specific content providers, such as proprietary dataset creators (Xie Yijiang, 2024). Practice has proved that the data-exclusive model not only improves licensing efficiency and professionalism, but also effectively avoids the emergence of "data oligopoly" and prevents the occurrence of monopolies (Han Yuxiao, 2025).

4. Determination of the content of the statutory licensing system

4.1 Package authorization mechanism of property rights

The infringement risk for the right of communication mainly appears in the output stage of generative artificial intelligence. When the content generated by the AI model is very similar to the original work, or even directly contains part of the original work, it may constitute an infringement of the right of reproduction and adaptation, especially when the content released through the network at the output stage may violate the right of information network communication. It should be noted that the training stage of generative artificial intelligence is closely related to the output stage. The former provides training materials for the latter through copying, format conversion, and other behaviors, and this process should be regarded as an auxiliary behavior of the output stage. If the content generated by AI in the output stage is substantially similar to the content used in the training stage, it may infringe the information network transmission right of the original work (Guo Dezhu & Zhang Yunwei, 2024).

4.2 Connection between copyright information labeling and artificial intelligence labeling

Many countries and regions have regarded the obligation to label content generated by artificial intelligence as a legal obligation. After labeling, right holders can clearly identify that the products are generated by artificial intelligence,

so as to take more targeted and effective rights protection measures, and thus better protect their rights. Therefore, the obligation to mark not only reflects the respect for the public's right to know, but also provides protection for the right holder (Yao Zhiwei, 2024).

In this context, creators of AI datasets and developers of generative AI should fulfill the identification obligation to clearly label the author and provenance of the work. This will not only help users understand the source of the generated content when using the work, but also help rights-holders better safeguard their rights and interests, and enhance the transparency of generative AI. The attribution and right distribution of AI-generated content can be regarded as a reasonable incentive mechanism, which helps to promote cooperation among technology developers, operators, and users, and promote the development of AI technology and applications. From the perspective of economic benefits, such rights distribution can help maximize the potential of artificial intelligence creation and promote the diversification and enrichment of cultural products. From the perspective of economic benefits, such rights distribution can maximize the potential of artificial intelligence creation, promote the diversification and innovation of cultural products, and thus improve the overall welfare of society. The legality and fairness of AI creation will become an important factor in stimulating more creativity and promoting industrial progress (Zhang Ping, 2024).

References

Chu, M. (2021). Artificial intelligence a challenge to the copyright infringement liability system and cope with. *Northern Legal Science*, (1), 138-150. https://doi.org/10.13893/j.cnki.BFFX.2021.01.013

Guo, D., & Zhang, Y. (2024). Emergent artificial intelligence training data infringement risk and legal response to. *Journal of Xiangtan University (Philosophy and Social Sciences Edition)*, *48*(5), 78-86. https://doi.org/10.13715/j.cnki.jxupss.2024.05.010

Han, Y. (2025). Copyright risk and solution of artificial intelligence large model training data. *China Publishing*, (2), 54-59.

Liu, X., & Xia, J. (2023). Function and practice of artificial intelligence labeling obligation. *China Foreign Trade*, (11), 51-53.

Neill, A., Thomas, J., & Lee, E. (2024). A framework for applying copyright law to the training of textual generative artificial intelligence. *Texas Intellectual Property Law Journal*, *32*, 225-250.

Wang, W. (2022). Challenges and responses to copyright limitation and exception rules by artificial intelligence. *Journal of Application of Law*, (11), 152-162.

Xie, Y. (2024). Copyright dispute and resolution of generative artificial intelligence works training. *China Editor*, (11), 38-46.

Yao, Z. (2024). Identification and prevention of copyright infringement of artificial intelligence products: Centered on the world's first case of generative AI service infringement judgment. *Local Legislation Research*, *9*(3), 1-17.

Zhang, L., & Wang, J. (2024). The dilemma of copyright in works of emergent use of artificial intelligence and relieve countermeasures. *Journal of Publishing*, (20), 75-80. https://doi.org/10.16491/j.cnki.cn45-1216

Zhang, P. (2024). Institutional problems and solutions of copyright legality of artificial intelligence-generated content. *Legal Science (Journal of Northwest University of Political Science and Law)*, (3), 18-31. https://doi.org/10.16290/j.cnki.1674-5205.2024.03.001