Managing Copyright Infringement Risks in Generative Artificial Intelligence Data Mining

Jing Li*

Department of Economics Law, East China University of Political Science and Law, Shanghai, China

* Corresponding Author Email: 212125010365@ecupl.edu.cn

Abstract. In the era of artificial intelligence, the rapid development of generative artificial intelligence represented by ChatGPT brings great convenience to human creation, but also causes many potential copyright risks and brings a series of new challenges to the field of intellectual property. The creative process of generative artificial intelligence mainly includes four stages: training data input, data learning, input instruction and content generation. Among them, the legal use of copyright works in the data input stage needs to be solved instantly. In order to solve this problem, we should first improve the "fair use system" under current Copyright Law, divide machine learning into expressive and non-expressive types according to whether generative artificial intelligence has expressive content output, and discuss whether it constitutes fair use separately. Secondly, in order to protect the legitimate rights of original copyright owners, it is also necessary to improve the transparency of training data and increase the "opt-out" mechanism. In addition, it is also necessary to clarify the tort liability subject of generative artificial intelligence by legislation when it does not constitute fair use.

Keywords: Copyright infringement, Rational use system, Opt-out mechanism, Tortious liability.

1. Introduction

With the continuous innovation of computer science, artificial intelligence has entered a new stage of development from assisting artificial intelligence to self-generated artificial intelligence. This stage affects every aspect of human creation, on the one hand, it brings great convenience to human creation, on the other hand, it also brings a series of new challenges to the existing copyright legislation and justice. At present, in judicial practice, there have been a large number of lawsuits against AI infringement, mainly concentrated in the field of AI painting and writing. For example, in the world's first AIGC infringement case "Stable Diffusion" [1], the U.S. photo library giant Getty Images filed a lawsuit against "Stability AI" for scraping millions of images online without permission to develop and train its AI image generation technology.

The case has attracted wide attention around the world. It has also sparked research on "fair use systems for generative artificial intelligence data mining". This paper aims to further analyze the types of copyright infringement that may be involved in the data input stage based on the operation mode of generative AI. Based on the study of existing domestic legislation and justice, combined with foreign experience, this paper aims to propose solutions to copyright risks that may be caused by generative artificial intelligence data mining. This lays a good legal groundwork for the future longevity of generative AI.

2. The Operation Mode of Generative AI and the Risk of Copyright Infringement at the Data Input Stage

2.1. Generative Artificial Intelligence Operation Mode

Generative artificial intelligence can independently generate pictures, text, animation, audio and other content by learning a large amount of training data to meet user needs based on user requirements. The specific operation mode includes four stages. The first stage involves inputting training data, which includes the acquisition and storage of training data, which includes the

acquisition and storage of training data that may fall within the copyright protection period and be used without the consent of the original copyright owner. The second stage involves data analysis. Generative artificial intelligence extracts patterns and forms a common expression template by analyzing a large amount of training data. The third stage is input instruction, which guides generative artificial intelligence in creating content that fulfills the user's needs. The fourth stage is content generation, which optimizes the expression model through learning from training data and constant adjustments by users, ultimately generating content indistinguishable from human works.

As mentioned earlier, it can be seen that generative AI cannot be developed without the use of data. High-quality and rich data is the basis of artificial intelligence training, and the size and quality of data directly affects the quality of the content generated. Artificial intelligence can form a creation model highly similar to human expression through the analysis and learning of a large amount of data, so that the artificial intelligence model has the ability to generate content autonomously in the face of different needs. However, the development of generative artificial intelligence is currently hindered by the lack of high quality data in Chinese. If we can solve the problem of insufficient data usage and make full use of China's high-quality works, it will effectively meet the data demand for the development of artificial intelligence [2]. However, at the same time, it may also seriously infringe on the rights of the original copyright owners.

In this regard, we should combine the legislative purpose of copyright law and the needs of social public interests, balance the interests of copyright owners and the interests of artificial intelligence development. Copyright law not only protects the generation of original content, but also promotes the use of works after their creation, in order to balance the exclusive rights of copyright owners with the interests of the public in using and enjoying works. Considering the interests of society and the development needs of the AI industry, copyrighted works can be conditionally opened to generative AI developers for use, while the interests of copyright owners need to be adequately safeguarded to achieve a balance of interests.

2.2. Risk of Copyright Infringement at the Data Input Stage

Based on the above modes of operation of generative artificial intelligence, it can be found that there is a substantial risk of copyright infringement in the operation of generative artificial intelligence. This is especially true for the training data input stage, as the nature of artificial intelligence creation is based on algorithms that require the conversion of a large number of works that are material carriers into digital works, and incorporate them into the database together with other digital works for learning. Through repeated training and adjustment, it can generate content that meets the needs of users in the output stage. In practice, the data input and learning of generative artificial intelligence are usually opaque and unpredictable, which is easy to infringe on the copyright rights of the authors work's data in practical applications. For example, in the data input stage, if the work within the copyright protection period is directly incorporated into the training data for learning without the authorization of the original copyright owner or without the fair use defense, it constitutes copyright infringement, which may specifically involve the right of reproduction and the right of adaptation [3]. And whether the use of copyrighted works in the data input stage is legal or not determines whether the subsequent output will constitute infringement. To address this issue, there are still many problems in China's legislation and justice that need to be improved. First, legislatively, the use of copyrighted works in machine learning does not be included in the twelve cases of fair use listed in Article 24 of the Copyright Law. Article 24 is a semi-closed legal provision. On the one hand, the first step of the "three-step test" is limited to "in the following cases", referring to the twelve specific cases listed in the article. On the other hand, the umbrella clause stipulates that other instance of fair use must be "stipulated by laws and administrative regulations". Therefore, by interpreting the general provision or the umbrella clause, machine learning cannot be recognized as fair use under Section 24 of the Copyright Act.

In addition, in July 2023, the newly promulgated "Interim Measures for the Administration of Generative Artificial Intelligence Services" (the "Interim Measures") also contains problems, such as

the interpretation of Article 7, which states that as long as it is a legitimate work and authorized by the original author, it can be included in the machine learning database. This article limits the use of machine learning materials to the "original author permission" model, and does not take into account the technological development costs brought by the acquisition of exponential training data for generative AI, as well as the algorithmic bias that may result from insufficient training data, which will further curb the development of generative AI.In essence, it over-protects the rights of original authors, while ignoring the needs of artificial intelligence development and the public interests of society. Secondly, Article 7 only addresses the protection of legal works, disregarding the utilization of illegal works by machine learning, which is still acknowledged as copyright in our country [4]. Moreover, Article 9 of the Interim Measures stipulates that providers should bear the responsibility of producers of network information content, representing an unwelcome implementation of the no-fault principle in the realm of generative artificial intelligence. This imposes stringent regulatory demands on artificial intelligence service providers, potentially resulting in excessive regulation and inadequate development incentives. It would be more prudent to ascertain tort liability based on the principle of fault-based liability, taking into account reasonable costs and accuracy, particularly when the industry is still in its nascent stage of development and the regulatory framework is not yet perfected [5].

Furthermore, in judicial practice, judges typically employ the "three-step test method" to assess whether it qualifies as "fair use", concentrating on analyzing its impact on the rights of copyright owners to establish its fair use status. Judges have great judicial discretion, and the judgment results are highly uncertain, so the vague use standard needs to be confirmed.

3. The Solution of Copyright Infringement Risk in Generative Artificial Intelligence Data Mining

3.1. Improving the "Fair Use" System

The development of generative artificial intelligence is inseparable from the learning of massive data, and the traditional licensing model requires the consent of all the authors of the works being learned, which is almost impossible to achieve. And the statutory licensing model will bring huge licensing fees, which will greatly increase the cost of artificial intelligence development, inhibit the enthusiasm of development investors, and thus hinder the development of artificial intelligence. Therefore, to address the legal use of copyright works in the input stage of generative artificial intelligence training data, the "fair use" system stipulated in article 24 of the current Copyright Law should be improved.

This paper suggests categorizing machine learning into "non-expressive" and "expressive" based on whether generative artificial intelligence produces expressive content output. It discusses the categorization of different types of machine learning [6]. The specific regulations are as follows: (1) Non-expressive machine learning refers to machine learning without expressive content output. For example, automatic identification technology copies, stores, organizes and transforms works but does not extract expressive content or have expressive content output. The use of such technology serves a "purpose conversion", meaning that it is not used for the purpose of disseminating the expressive content of the original work to the audience [7]. This type of use is considered non-work use and does not constitute infringement. Therefore, it is not liable for infringement. (2) Expressive machine learning involves the use of works in the sense of copyright law and is initially recognized as infringement. Specifically, it can be divided into the following three situations: Firstly, it involves learning public expression, where machine learning imitates public expression and utilizes works from various authors. The extracted elements of expression typically do not replicate the copyrighted original expression of the works being learned. Instead, the approach involves "seeking common expressions while reserving differences", understanding the basic compositional patterns of the works to create a general template and generate new expressions through "content conversion" [8]. This type of use falls within the scope of copyright law but is considered fair use, exempting it from

infringement liability. It is important to note that if the artificial intelligence lacks sufficient "intelligent", resulting in generated content that closely mirrors the expression of the learned work, creating a "substantial similarity", it may not quality as fair use and could incur corresponding infringement liability. However, an artificial intelligence advances, the likehood of such a scenario diminishes over time until it eventually disappears. Secondly, it is to learn the specific author and generating a personalized expression of the author ,which will inevitably have a substitution effect on the author's work and harm its potential market, thus harming the rights of the original copyright owner. This situation is difficult to constitute a fair use, should bear tort liability. Thirdly, it is learning for non-profit scientific research activities, which should be recognized as research-based fair use. Research-based fair use can be expanded interpreted as individual research (1) and school scientific research (6) in Article 24 of the Copyright Law of China as research-based fair use and exempt from tort liability [9].

3.2. Increasing Transparency of Generative AI Training Data and Add an "Opt-out" Mechanism

The "fair use system" is the result of balancing the rights of the original copyright owners and the public interests of society. It should not blindly sacrifice the interests of the original owners. Therefore, legislation should stipulate that AI technology developers or service providers must disclose the sources of copyrighted works used to train their AI. This measure aims to protect the authors' right to know and respect their creative results, promoting knowledge dissemination and encouraging creation. Additionally, the generated content should be marked with digital watermarking or similar methods to prevent users from registering artificial intelligence-generated content for copyright, which could further infringe upon the rights of the original copyright owners.

In addition, even if the use of works in the training data input stage constitutes fair use to obtain a legitimate evaluation, the use of works in the content generation stage is difficult to obtain the legitimate basis. In particular, based on the legislative purpose and spirit of the copyright Law, the content generated by generative artificial intelligence does not constitute a work [10], which should not be protected by copyright and directly enter the public domain and be widely disseminated. This will inevitably have a substantial impact on the real or potential market of the works being studied. Therefore, in order to protect the rights of copyright owners, we should also learn the cautious scheme of "limited opening" in the EU. That is, to set up an "opt-out" mechanism in addition to the fair use system. To be specific, it is considered that copyright owners agree to the use of their copyrighted works, but give them the option to refuse the inclusion of their works in the training data [11], so as to take into account the interests of copyright owners, ease their opposition, and increase the acceptances of the "fair use system".

Of course, the opt-out of fair use does not necessarily mean that the work is completely excluded from the training dataset. Technology developers and service providers can still choose to keep their works in or reintroduce them into training dataset through licensing agreements by paying corresponding copyright fees to the original authors of the works. A similar practice is already taken in the European Union, with Spawning, a company worked on building an "AI consent mechanism". On the one hand, it provides the artist with a query service of "whether the work is used for artificial intelligence training", helping them to opt out from other people's training datasets. On the other hand, it provides copyright compliance services for artificial intelligence companies and help them to obtain copyright licenses.

3.3. Clarify the Subject of Tort Liability

The infringement of works by generative AI becomes explicit through the generation of content. Based on the previous discussion that non-expressive use does not amount to the use of works and is not subject to tort liability, further infringement analysis will not be conducted here. However, in the case of expressive use, the situation differs. When the generated content from generative artificial intelligence exhibits original expression during the output stage, it should be categorized as a "content

conversion", qualifying as fair use and therefore exempt from tort liability. On the other hand, if the generated content merely replicates or rearranges the learned work without adding significant originality, resulting in substantial similarity to the original work, it should not be considered fair use but rather as copyright infringement [12]. In practice, when assessing infringement considering the input of previous works, the determination may follow the standard of "contact + substantial similarity". In other words, if the generated content is substantially similar to the original work and lack prior permission from the copyright owner, the infringement of the right to copy can be established directly without the need to evaluate the degree of similarity in terms of quantity and quality. The consideration of similarity in quantity and quality comes into play when the defendant subsequently claims "fair use" as a defense.

The resulting tort liability should be clearly defined in legislation to protect the rights of the original owners. This paper argues that the tort liability should be categorized and discussed based on different development models.

When the service provider purchases the model developed by the technology developer and provides it to the user directly, it should be established that the technology developer bears the corresponding copyright infringement liability. Because it understands and can control the data input of generative AI. Unless the developer can prove that the model was not exposed to the original work during the data training process. As a result, the problem of "algorithmic black box" can be effectively solved by comprehensively checking the algorithms, training data, and output content, so as to better improve the interpretability of artificial intelligence.

When the service provider fine-tunes the model to meet the requirements of specific application scenarios, both the pre-training in the technology development stage and the fine-tuning before the later market application will essentially shape the generative AI algorithm model. Take ChatGPT for example, in the pre-training stage, technical developers use massive work data to make the machine form a common expression template. In the fine-tuning stage, service providers use high-quality code fine-tuning corpus and conversation fine-tuning corpus to adjust the basic model. At this time, based on the general principles of tort law, the technology developer and the service provider should be respectively liable for the infringement content generated. However, in individual cases, it is difficult for rights holders to clarify the boundaries of the respective capability of the technology supporters and service providers. Therefore, for the sake of the efficiency of rights protection, the appropriate solution is to ask the service provider to take measures to optimize the algorithm model. As for how to achieve the optimization goal, it is left to technical developers and service providers to resolve through the internal agreement [11].

When the user gives wrong instructions, that is, actively input the copyrighted work, guide the generative artificial intelligence to generate content being "substantially similar" to the learned work, and damage the rights of original copyright owners, the user shall bear the infringement liability to promote the reasonable distribution of infringement liability [2].

4. Conclusion

Through the above research on the operation mode of generative artificial intelligence, we have clarified the great significance of data to the development of artificial intelligence. And the copyright infringement problems caused by the use of copyright works in the data input stage are emphatically analyzed to ensure the legitimacy of the subsequent content output. In view of the problems in domestic laws and regulations such as Article 24 of the "Copyright Law" and Articles 7 and 9 of the "Interim Measures for the Management of Generative Artificial Intelligence Services", as well as the problems existing in judicial practice, the current "fair use system" should be improved, that is, machine learning should be categorized and regulated as follows. (1) Non-expressive machine learning belongs to non-work use, does not constitute infringement, does not have tort liability, and does not need to be legalized through fair use. (2) Expressive machine learning is subdivided into three categories: Firstly, mass machine learning constitutes fair use and is exempt from infringement

liability. Secondly, individual learning, which is difficult to constitute fair use, should bear tort liability. Thirdly, non-profit use, constitutes research-based fair use and is exempt from tort liability. At the same time, in order to protect the rights of authors and balance the interests of rights holders and the public interest of society, generative AI technology developers or service providers should be required to disclose their copyright work training data, and add an "opt-out mechanism" to allow rights holders to refuse to have their works included in the training data. Finally, we should clarify the liability for infringement in cases that do not constitute fair use to further protect the rights of copyright owners and increase the acceptability of the "fair use" system. However, the above research remains at the theoretical level and needs to be implemented in legislation and judicial practice in the future, so as to promote the healthy development of generative artificial intelligence.

References

[1] Getty Images(US), Inc.v. Stability AI, Inc., Feb.3, 2023, No.23-cv-00135.

[2] Gu Nanfei, Fang Zhouzhi. Reasonable boundary and infringement regulation of generative AI use works such as ChatGPT. Digital Library Forum, 2019,19 (07): 1-8.

[3] Song Weifeng. Generative AI communication paradigm: Copyright risk and regulatory construction of AI-Generated content -- Based on the world's first AIGC infringement Case. Press & News, 2023, (10): 87-96.

[4] Zhan Ailan, Tian Yinong. Copyright risk in generative artificial intelligence machine learning and its solution path. Electronic Intellectual Property, 2023, (11): 4-14.

[5] Xu Wei. On the legal status and liability of generative artificial intelligence service providers: A case study of ChatGPT. Law Science (Journal of Northwest University of Political Science and Law), 2023, 41 (4): 69-80.

[6] Benjamin L., W. Sobel. Artificial intelligence's fair use crisis. Columbia Journal of Law & the Arts, 2017, 41: 45-49.

[7] Matthew Sag. Copyright and copy-reliant technology. Northwestern University Law Review, 2009, 103(4): 1983.

[8] Wu Handong. Questions on the copyright law of artificial intelligence-generated. Chinese and Foreign Law, 2020, 32 (03): 653-673.

[9] Lee An. Copyright rules for machine learning: Historical implications and contemporary solutions. Global Law Review, 2023, 45 (06): 97-113.

[10] Wang Qian. On the characterization of artificial intelligence-generated content in Copyright Law. Forum on Political Science and Law, 2023, 41 (04): 16-33.

[11] Shao Honghong. Research on copyright infringement of generative artificial intelligence. Publication and Distribution Research, 2023, (06): 29-38.

[12] Peng Feirong. On the copyright infringement risk of data in algorithm creation and its resolution. Journal of Application of Law, 2023, (4) : 46-55.