A Study of Fair Use in Training AI Corpora

Zheyu Pei
Law School, Zhejiang Gongshang University, Hangzhou, Zhejiang, China

Keywords: AI Corpus Training, Fair Use, Copyright Law, Generative AI Program.

Abstract: In the digital age, artificial intelligence technology is developing rapidly. The emergence and extensive application of generative artificial intelligence have drawn scholars' attention to the legal issues in the training of artificial intelligence corpora. This paper summarizes and analyzes the qualitative problems of data mining in AI corpus training by domestic and foreign scholars, and focuses on the legitimacy and necessity of fair use. Through the re-search on the identification of fair use of AI corpus training in copyright law, this paper points out that fair use system can better balance the rights and interests of creators and public interests and promote the development of new technologies. In view of the limitations of the enumerated provisions of China's Copyright Law, the study suggests that AI training should be explicitly included in the scope of fair use, and that the EU should be used as a reference to establish special exception clauses, as well as to establish a mechanism for compensating authors in commercial use scenarios, so as to achieve a dynamic balance between promoting techno-logical innovation and protecting the interests of creators.

1 INTRODUCTION

In the current digital environment, artificial intelligence has gradually come into the public's view, and has caused a non-negligible impact on all fields of human society. For example, the emergence and application of generative AI bring transformative potential to the field of education, reforming educational tools as well as promoting the development of critical thinking (Grant, 2023). The emergence of generative AI has also boosted the healthcare industry with the ability to gather large amounts of information to facilitate the refinement of treatment protocols (Moulaei, et al, 2024).

Similarly, the impact of AI makes a difference to the field of intellectual property, as a new technology worthy of research, its emergence brings brand new opportunities for the development of intellectual property rights globally, as well as new dilemmas for the protection of intellectual property rights. In terms of copyright, the use of generative AI can be focused on two processes, the data input phase and the generator output phase. Regarding the data input stage, the AI program will absorb a large number of copyrighted works for training. The legality of using copyrighted materials for artificial intelligence model training and whether such practices constitute copyright infringement currently lacks specific legal regulations and standardized guidelines. Academic circles hold divergent perspectives on this issue: some scholars contend that utilizing copyrighted works without explicit authorization for model training does not violate intellectual property rights, while others maintain opposing viewpoints. This study, therefore, focuses on the research of fair use in AI corpus training, synthesizing existing academic perspectives. Building upon this foundation, the paper proposes a reformed framework for reasonable application, aiming to facilitate the synchronous development between China's fair use system and artificial intelligence technologies.

2 QUALITATIVE DOCTRINE FOR DATA MINING DURING AI TRAINING

The process of collecting and processing large amounts of training data and searching and analyzing the information hidden in it through algorithms during artificial intelligence training is known as data mining. Scholars in academia hold divergent positions regarding the legal characterization of data mining. A segment of academics asserts the justifiability of employing copyrighted works for AI

Pei, Z.
A Study of Fair Use in Training AI Corpora.
DOI: 10.5220/0014381400004859

In *Proceedings of the 1st International Conference on Politics, Law, and Social Science (ICPLSS 2025)*, pages 381-386
ISBN: 978-989-758-785-6

training under specific circumstances, while opposing scholars categorize such utilization as prima facie copyright infringement.

Scholars who support the infringement theory argue that the training of large models of artificial intelligence belongs to commercial use, and is not within the fair use circumstances categorized by the current copyright law. Therefore, it should be recognized as a copyright infringement (Yi, 2024). Such scholars adhere to the statutory principle, but will appear to lack practicality in the light of evolving AI technology.

On the other hand, the scholars who advocate the justification of data mining have further posited distinct doctrinal frameworks, developing novel theoretical constructs within the parameters of technological neutrality principles. Part of them have put forward the theory of non-expressive use, which refers to the use of an original work for the purpose of treating it as factual information and making functional use of it, rather than utilizing the original expression in the original work or reflecting the artistic value of the original work in the new work (Jiao, 2022). Scholars holding this view argue that AI training constitutes the analysis and synthesis of patterns from large-scale data, rather than utilization of the original work itself, thus being a non-expressive use that does not amount to copyright infringement; Some scholars adhere to the theory of transient reproduction, claiming that copying under copyright law constitutes preparatory acts for dissemination, and only when the duplicated material is utilized for distribution does it infringe the right of reproduction (Tu, 2024). The purpose of the copying behavior in AI corpus training is for model training, and is only used to generalize and analyze the logical patterns behind the content, which is a temporary use of the data, and will not be disseminated to the outside world, so this temporary copying behavior does not constitute infringement under copyright law; Besides, other scholars advocate the fair use doctrine, maintaining that AI training adheres to the value principle of technological neutrality. Based on the public benefit and fundamental nature of AI models, they assert the substantive utilization of copyrighted works during machine learning possesses legitimate purpose and does not adversely affect normal exploitation of the original works. Thus, this kind of usage is permissible to be categorized as fair use (Yi, 2024).

3 FAIR USE IN ARTIFICIAL INTELLIGENCE CORPUS TRAINING

The legal characterization of data mining in AI corpus training remains contentious. This paper posits that the fair use doctrine proves more congruent with China's judicial reality and global trends, effectively facilitating technological advancement while stimulating socioeconomic development. In contrast, the temporary copying doctrine focuses on the transient and technical nature of the copying behavior, such as temporary copying in caching or browsing, and determines that the use of another's work in AI training does not constitute infringement in terms of the duration and technical characteristics of the copying. The non-expressive use doctrine, on the other hand, emphasizes the use of a work that does not involve expressive content, such as data analytics or text mining, and focuses on the fact that the purpose of the use is not to convey information about the work itself. The fair use doctrine upholds the value principle of technological neutrality. Through an interest-balancing framework, it comprehensively evaluates factors including the purpose, nature, and quantity of copyrighted work utilization during training processes, along with market impact. By determining the conversion ratio of public interest in the usage behavior, it concludes such practices fall outside the scope of infringing acts. This more flexible approach to adjudication can accommodate the rapidly evolving information technology and the commercial interests behind it. Thus, through the above comparative analysis, the author is more inclined to analyze the use of others' works in AI corpus training from the perspective of fair use.

3.1 Rationality

The fair use system is fundamentally about limiting individual copyrights and protecting the public interest, and its underlying logic is a balance of interests, preventing the boundless expansion of authors' copyrights while facilitating technological development as a means of expanding the public interest in society. In the absence of a fair use system, authors' copyright over their works is too broad, which is not only unfavorable for management, but also unfavorable for technological advances in the public domain, as authors' blocking of their works can negatively affect the efficiency of overall development. Chinese scholar Xu Xiaoben's view that AI technology belongs to the universal

technology that can benefit human society and promote scientific and technological progress, and therefore should be included in the fair use system according to the pursuit of the public interest. And he believes that the use of works in the training of AI will not affect the normal use of the original works, and that as long as it is complemented with other mechanisms of copyright protection, it can also take account of the original author's and the interests of the legislation (Xu, 2024).

The fair use system was initially dominated by the four-factor determination rule. The 19th-century case Folsom v. Marsh is considered to be the first case on fair use in the U.S. In this case, Judge Joseph Story first proposed a four-factor analytical framework that would become the foundational standard for assessing fair use (Justia, 1841). This analytical framework focuses on four key considerations: the purpose and character of the use; the nature of the original copyrighted work; the amount and substantiality of the portion used; and the effect of the use on the potential market or value of the work.

Then in the 20th century Campbell case, judges introduced the "transformative use" criterion, elevating the fair use doctrine from its traditional four-factor assessment framework into a new phase of development. (Justia,1994) Wang Qian, a scholar in China, believes that "transformative use" refers to the use of such works by adding new content and ideas to the original work, so as to make the original work show a new value. It is not a simple copy of the content of the original work, but an approach of realizing the conversion of the work's function or purpose (Wang, 2021). Also, Chinese scholar Wu Handong points out that within the criteria for determining fair use, the rule of "transformative use" emphasizes whether the new work's utilization of the original material demonstrates transformative purpose and character, rather than rigidly adhering to restrictions on the nature or quantity of the original work's usage (Wu, 2020). The perspectives of the aforementioned two parties focus more on the transform of subjective intentions when using the original work, creating new value based on the factual content of the original work, which is in perfect alignment with the essence of artificial intelligence's corpus training. Artificial intelligence corpus training is not a direct use of the original work, but through data analysis to summarize the patterns behind the original work, so as to form a logical generation mode and generate new content for the user. That's the reason why training method can constitute the so-called "transformative use" provisions.

Furthermore, Chinese scholar Yi Jiming has proposed a special version of "transformative use" criterion. Within this framework, he emphasizes that such usage not only entails a subjective shift in purpose but also involves objective technological innovation, thereby achieving a complete transformation and value-added enhancement of the original work's significance (Yi, 2024). This evaluative approach can help trainers and users pay more attention to whether the utilization of a work contributes to the advancement of science and technology as well as the realization of societal public interests, thereby fostering the creation and value enhancement of new technologies. According to that, Artificial intelligence, a new technology which can promote scientific and technological progress, absolutely has rationality for its training process. What's more, the author contends that traditional "transformative use" and the new version of it do not need to form successive phases but should coexist to accommodate the widespread implementation of artificial intelligence. For instance, the criteria for traditional one could govern the training of most generative AI systems in producing images, texts, and similar outputs, while the new standard might apply to the development of AI models centered on technological breakthroughs.

Of course, there are different views in the academic community about whether AI corpus training can constitute fair use. Chinese scholar Wang Xuelei acknowledges that the utilization of works during artificial intelligence training constitutes an exploration and analysis of the patterns underlying the works. However, she argues that the object of such use is the data extracted from original works rather than the works themselves, thereby eliminating the necessity for copyright regulation in this context (Wang, 2025). This study, however, contends that the data utilized for analysis originates from copyrighted works rather than existing in isolation. Such works should not be categorically excluded during analytical processes. Furthermore, from the perspective of balancing interests, reducing works to mere data vessels through disaggregation undermines the rights and benefits of human authors. Meanwhile, it must be emphasized that the fair use system inherently functions as a product of balancing competing interests. We should keep a balance between the interest of individual copyright and public interest of the whole society. Some scholars, motivated by concerns for copyright holders' interests, argue that categorizing works created by human intellectual labor as training materials for artificial intelligence under fair use provisions would

lead to the systematic disregard of human creators' rights and contributions (Zou etal, 2024). This paper won't deny the restriction of individual copyright could be brought from the fair use system. Nevertheless, excessive emphasis on protecting individual authors' interests could impede technological progress and generate adverse effects on the advancement of public interests. Facing this problem, people should actively seek or establish a new system of benefit distribution that can provide protection for human authors, rather than just preventing the realization of the fair use in AI corpus training.

Above all, through the analysis of diverse scholarly viewpoints of fair use and some different cases, this research concludes that data mining within artificial intelligence training can be demonstrated as a form of fair use with both possibility and rationality.

3.2 Necessity

From the points of modern technology, the training of AI relies on a large scale of data, arranging from text, picture, video in diverse field. Therefore, if the usage of every single piece of data requires separating authorization, the cost will be apparently increased and the development of technology iteration will be hindered.

From the points of commercial interest, enterprises will face substantial compliance costs and legal risks if too many authorizations are in need. In contrast, the system of fair use can significantly reduce upfront investment, encourage more companies to invest in the AI track, and promote overall innovation in the industry.

In addition, the application of fair use in AI corpus training is not a regional issue, but a global one. With ChatGPT, DeepSeek, and many other generative AI programs being put to use around the world, a large number of human authors' works are being dismantled and learned by the training of AI. In the face of this common problem, it has become an international trend to categorize AI corpus training as fair use. Of course, different countries do not adhere to a unified framework of fair use determinations. According to their different cultural backgrounds and histories of jurisdictions, the appropriate legal system has been established in accordance with the specific practice of each country. For example, the United States has taken a liberal stance based on "transformative use", and case law in the United States has brought text data storage and mining within the scope of fair use, based on the doctrine of "transformative use" (Wu, 2020). What's more, articles 3 and 4 of the Copyright in the Digital Single Market Directive promulgated by the European Commission establish exceptions for "text and data mining for scientific research purposes" and "exceptions or limitations for text and data mining" respectively. These provisions constitute mandatory exceptions, requiring all member states to implement TDM exception rules within their domestic legal frameworks as minimum standards (Bao, et al, 2025). Both the U.S. and EU regulations have already established the reasonableness of data mining by law, and this change with the times further signals the development of AI technology and the urgent need to regulate the legal issues behind it.

Based on the above analysis, it can be concluded that it is reasonable and necessary to classify the use of others' works in AI corpus training as fair use. Of course, from a human creator's perspective, having his or her work used for AI corpus training puts him or her at risk. The use of a human author's work by a machine is out of the control of the original author and does not generate revenue for the author, thus creating an imbalance of interest between the AI and the human author (Wu, 2020). To be specific, because of the nature of AI training, a large amount of work is put into it, and the original authors of the work are most likely unaware that their work is being used to train the AI, let alone make a profit from it. And the process of AI corpus training is massive, fragmented, and cut up, so in case of infringement disputes, the original authors will face difficulties in pursuing their responsibilities. Therefore, under the present environment of the digital age and technological development, we should adhere to the basic principle of technological neutrality as well as the pursuit of commercial interests and technological advancement while taking into account the interests of human authors, so as to reform the original system and find a fulcrum for balancing interests. That's why this research believes that a balanced benefit-sharing mechanism for creators should be established while applying AI corpus training to reasonable use, taking into account commercial interests and technological development.

4 DEFICIENCIES IN THE LAW AND RECOMMENDATIONS

4.1 Deficiencies

Although applying the fair use system to AI corpus training is conducive to protecting the public interest

and balancing conflicting interests, the current legal provisions on fair use in China are rather thin. Currently, there are only the 12 circumstances listed in Article 24 of the Copyright Law and the "other circumstances stipulated by laws and administrative regulations", as well as the three-step standard of "limited to certain special circumstances" introduced by the Regulations for the Implementation of the Copyright Law. China's judicial practice regarding the determination of fair use is still based on specific enumeration, which means that judge does not have the right to create new exceptions of fair use. In that case, judge does not have the right to directly expand the interpretation of fair use to the AI corpus training data mining as well.

In addition, AI corpus training cannot be interpreted by the existing twelve circumstances in the law. The use of works in AI corpus training is usually large in scale and quantity, which is not in line with the requirements of "appropriate citation". Besides, the training of AI basically belongs to the behavior of enterprises, and most of the training is carried out for commercial purposes, which is not in line with the purposes of "personal learning", "news reporting", "education and research" or "national public service".

The existing law does not specify whether the emerging AI corpus training can be categorized as fair use, which cannot meet the rapid development of the existing AI technology and the status quo of the massive use of AI technology. For the sake of the relevant interests and protections of the creators, the fair use system in China should be reformed.

4.2 Reinventing Fair Use System in the Age of Artificial Intelligence

Regarding the difficulty of using the existing fair use system for data mining under the current legal environment, there are different views in academic field. For example, Chinese scholar Huang Xijiang suggests that even if there is no special provision for fair use of "text and data mining", the fair use in AI training can still be regulated by the general provision in Article 24 of the Copyright Law (Huang, 2024). On the contrary, Chinese scholar Lin Xiuqin proposes that we should learn from the U.S. model of "fair use plus specific enumeration", which stipulates the factors for fair use based on four-factor determination rule, and then enumerates the common ways of fair use (Lin, 2021). What's more, Chinese scholar Wu Handong contends that "text data mining" should be added to the list as a new case of fair use (Wu, 2020).

Whether it is to explicitly list AI corpus training as a special case of fair use or to recognize it through judicial interpretation as a case stipulated in the underpinning provision of Article 24 of the Copyright Law, the relevant legal system should be in line with the rapid development of the trend of AI technology to make certain changes. The author believes that, in contrast, the addition of AI corpus training to become a specific case of fair use can be more intuitive and quick to regulate the relevant legal issues, but it should be differentiated between the use of AI corpus training. For an instance, if it is used for scientific research and non-commercial purposes, such as teaching, then it will be directly applicable to the fair use. And if it is used for commercial purposes, it should be established with the author of the original work to pay the appropriate compensation.

5 CONCLUSION

To summarize, the rapid progress of modern science and technology has prompted the human society to face the potential legal problems of AI technology. The society have to commit that the contradiction between the wildly use of generative AI for other people's works and the limited provisions of the existing law needs to be solved urgently. This paper believes that the fair use theory can achieve a smooth and fair state between the copyright of human authors, the commercial interests of AI training enterprises and the interests of the overall development of society. Therefore, in order to promote the development of the technology and improve the efficiency of the judiciary in the face of this problem, the legislator should clarify the specific circumstances of fair use in the legal revision, and incorporate the AI corpus training into the scope of the fair use, and establish a unified standard of this issue. At the same time, establishing a mechanism for the original author of the work with the payment of appropriate compensation is a good way to balance the commercial interests of the AI training enterprises with the human author's copyright, so that people can promote the upward development of society.

REFERENCES

Bao, S. & Xiao, D. 2025. The copyright law consequences of generative artificial intelligence training data: An analysis of the EU copyright exception rules and their implications for China. Library Forum.

Cooper, G. 2023. Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. J Sci Educ Technol 32: 444-452.
Huang, X. 2024. A peek into China's copyright law adherence and change in the age of intelligence by ChatGPT. Intellectual Property 8: 124-125.
Jiao, H. 2022. The copyright risks of data acquisition and utilization in the creation of artificial intelligence and the path to resolving them. Contemporary Law 4: 130.
Justia U.S. Supreme Court. 1994. Campbell v. Acuff-Rose Music, Inc. 510 U.S. 569. 2025.4.1. https://supreme.justia.com/cases/federal/us/510/569/
Justia U.S. Supreme Court. 1841. Folsom v. Marsh. 9 F.Cas. 342 (C.C.D. Mass. 1841). 2025.4.1. https://law.justia.com/cases/federal/district-courts/massachusetts/madce/9fcas342/4104271/220/no-4.html
Lin, X. 2021. The reshaping of the copyright fair use system in the age of artificial intelligence. Legal Studies 6(6): 182-183.
Moulaei, K., Yadegari, A., Baharestani, M., et al. 2024. Generative artificial intelligence in healthcare: A scoping review on benefits, challenges and applications. International Journal of Medical Informatics: 105474.
Tu, T. 2024. Determination of copyright infringement of machine learning: Beyond the theory of non-expressive use. Politics and Law 10: 166-168.
Wang, Q. 2021. A course in intellectual property law (7th ed.). People's University of China Press: 330.
Wang, X. 2025. Reflections on the application of copyright fair use system for artificial intelligence data mining. Hebei Law Journal 3: 190-191.
Wu, H. 2020. The question of copyright law for works generated by artificial intelligence. Zhongwai Jurisprudence 3(3): 659-661.
Xu, X. 2024. The copyright fair use of artificial intelligence model training under the perspective of technology neutrality. Law Review 4: 94.
Yi, J. 2024. Research on the reasonable use of large model corpus training. China Copyright 6: 7-8, 19.
Zou, H., Bi, M., Pu, J., Zhao, L., & Yan, L. 2024. Exploring the compliance of artificial intelligence training data. Modern Commerce and Trade Industry 19: 33.