

The Comparative Analysis of Fair Use of Works in Machine Learning

Yunlan Zhu^{1,2,*}

¹ Shenyang University of Technology, Shenyang, China

² School of Law, Sapienza University of Rome, Rome, Italy.

Abstract. Before generative AI outputs the content, it copies a large amount of text content. This process is machine learning. For the development of artificial intelligence technology and cultural prosperity, many countries have included machine learning within the scope of fair use. However, China's copyright law currently does not legislate the fair use of machine learning works. This paper will construct a Chinese model of fair use of machine learning works through comparative analysis of the legislation of other countries. This is a fair use model that balances the flexibility of the United States with the rigor of the European Union.

1 Using works in Machine Learning: principles and legal logic

1.1 How Machine Learning works

Machine Learning is a subfield of artificial intelligence that involves automatically detecting meaningful patterns in data and using the detected patterns for certain tasks. Briefly, the learning process involves an algorithm that takes training data (representing past knowledge or experience) as input and output information that other algorithms can use to perform tasks such as prediction or decision making [1].

For example, ChatGPT, the world's popular chat robot, is a large language model among generative machine learning models. In ChatGPT, a large amount of text is first pre-trained to generate a general model [2]. When the user inputs a question or information, ChatGPT will generate a corresponding answer or expression based on the previously learned language model.

1.2 Copyright involved in Machine Learning works

The development and application of machine learning relies on the "feeding" of massive information including copyrighted works, which specifically includes multiple behaviors such as extracting, copying, and format transcoding of works [3]. The process is likely to fall into the regulatory scope of the right of reproduction.

In order to carry out machine learning, staff need to translate, label, organize, and summarize the data. These behaviors may infringe the derivative right. In order to conduct machine learning or to achieve the verifiability of

research results, relevant personnel need to transmit data or text through the Internet, or upload it to the "cloud", etc. From the perspective of copyright law, it may infringe the right of communication to the public [4].

There are three copyright rights involved in the use of machine learning works. So, as artificial intelligence creation using machine learning technology becomes more and more common, it not only brings prosperity to cultural creation, but also brings challenges to the traditional copyright system [5].

2 Fair use of works is necessary for machine learning in China

2.1 The current fair use rule lacks usability

The fair use clause in China's Copyright Law has listed twelve specific situations that fall under fair use, but it is not clear which situation the use of works by machine learning falls under.

On this basis, we can seek legal interpretation to determine whether the use of works by machine learning can be covered by the circumstances clearly enumerated in Article 24 [6]. However, in fact, these situations all have applicability issues to a greater or lesser extent and cannot be applied to the utilization of machine learning works involved in the development of artificial intelligence applications. The first paragraph limits the subject and can only be used by individuals. Paragraph 6 limits the quantity of use, but the use of machine learning in works is large and extensive. Paragraph 8 limits the usage, which can only be displayed or preserved.

Although China's new "Copyright Law" in 2020 has added the general provisions of "other circumstances

* Corresponding author: elainejoo@126.com

stipulated by laws and administrative regulations" in the terms of Fair Use, the specific circumstances of my country's copyright Fair Use are stipulated in the "Copyright Law" and "Information Network Communication". In addition, there are no other laws or administrative regulations that stipulate the Fair Use system, and the behavior of using works by Machine Learning is difficult to be included in the scope of these situations. Therefore, when laws and administrative regulations do not provide for other circumstances, the catch-all clause cannot play a catch-all role, nor can it solve the dilemma of applying the Fair Use system to Machine Learning.

2.2 Rationality of Machine Learning fair use

After the authors create the works, they are copyright owner who can get interests from works. As a kind of private right, copyright law mainly focuses on the protection of private right. Others, such as the algorithm company and platform users, represent the public interest in public domain. However, Machine Learning may break the balance between the intellectual property and public interest [7]. Under the guidance of the public interest, copyright law delineates the scope of the public domain, so that works within this scope can be used freely and without compensation.

For example, the fair use system in copyright law. The purpose of fair use is to limit the property rights of copyright and reduce the scope of economic rights of the copyright owner. In this case, the copyright owner loses the ability to restrict the behavior of certain users and reasonably transfers some of its rights and interests [8]. Because of the dynamic balance between exclusive rights and the public domain in IP law, the fair use system has become a very important legal system for economic and social development and technological innovation in contemporary world.

3 Representative legislative models of machine learning fair use of works

3.1 United States: Application of the transformative use rule

Although machine learning is technically further and more complex than text data mining, the two are closely related. If text data mining is a copyright exception, then this copyright exception applies equally to machine learning [9]. Currently, statutory laws and case law in many countries have established copyright exceptions for text data mining.

U.S. jurisprudence has established an "unconditional exception" model that does not set any conditions for use of copyright exceptions for text mining, and has established a "transformative use" rule.

The U.S. copyright law adopted open legislation with "four elements" as its core in 1976. U.S. courts flexibly weigh the proportion of each factor in the judgment of a specific case, thereby expanding the application space of

the fair use system in practice.

Google launched the Google Library project in 2004, which unites the world's major research libraries and collects a large number of book resources, converting these books into digital texts. However, the Writers Guild of America took Google to court in 2005, arguing that the book resources scanned and uploaded by Google contained many works that were used without the authorization of the original copyright holders, which was a copyright infringement [10].

In 2013, the U.S. court ruled that Google's actions did not constitute copyright infringement. According to the "four elements of fair use" stipulated in Article 107 of the U.S. Copyright Act, the Google Library Project's actions had a "transformative purpose." In 2016, the Writers Association appealed to the Federal Supreme Court, which ultimately upheld the original verdict [11].

However, the recent "TV Eyes case" made a judgment different from the Google case, which to a certain extent reflects the court's rigorous attitude towards judging the Fair Use of Machine Learning. The defendant in this case, TV Eye, is a media company. The company records the programs broadcast by various TV channels, digitally classifies the recorded content, and provides clips of no more than 10 minutes long to customers, so that customers can locate the programs they want according to their preferences.

The completely different judgment results of the Google Books case and the TV Eyes case show the attitude of the US courts on the Fair Use of Machine Learning. That is, the recognition of the transformational use of Machine Learning works constitutes Fair Use, but the presentation of the output results needs to be considered independently. If the effect of public communication on copyright content has the possibility of replacing the original work and occupying the market of the copyright owner, it will be deemed as an infringing use.

3.2 Japan, the United Kingdom: independent and clear exceptions

3.2.1 Japan

In 2009, Japan's "Copyright Law" added a new exception to copyright law - "Copying for the purpose of analyzing information", which is considered to be the earliest legislative example involving the provision of copyright exceptions for the use of artificial intelligence works. However, this computer exception is limited to the scope of "information analysis", and it is difficult to cover new ways of using works under the background of Machine Learning and deep learning technology.

The Japanese government has also pointed out that the provisions of the old law are not enough to meet the current development status and long-term needs of the artificial intelligence industry. Although this provision only applies to works and does not apply to databases, in general, compared with other countries. Japan's copyright legislation has a system of restrictions on the exercise of copyright that gives considerable freedom to text and data mining [12].

Japan's "Copyright Law" revised in 2009 added new restrictions to the right of reproduction, that is, when using a computer for information analysis, the user has the right to cache or adapt the work within the necessary limits (including Cache of Derived Works). Japan's Intellectual Property Protection Organization (AIPPI) pointed out in a report on "Exceptions to Copyright Protection and Licensing of Works in a High-Tech Digital Environment" that the type of work does not affect the application of this clause. However, at the same time, this article stipulates that the use of databases does not apply to this rule, which is also consistent with the "three-step test method" stipulated in the "Berne Convention", that is, the laws of member states have the right to allow the reproduction of the above-mentioned works under certain special circumstances. As long as this copying does not damage the normal use of the work or unreasonably harm the legitimate interests of the author.

3.2.2 UK

The United Kingdom revised its copyright law in 2014, and made changes to copyright law based on the recommendations in the famous "Hargreaves Report" to adapt to the development of the times, including the copyright system for the Fair Use of text and data mining [13]. The "Hargreaves Report" believes that it is necessary to establish a Fair Use copyright system for text and data mining, because this technology is very practical and representative in the new era, and cannot be restricted by copyright laws and cannot play its role.

Unlike the "unconditional exception" in the US, the text and data mining exception in the UK is a "conditional exception". In the main element, purpose element and form element, there is a certain degree of restriction. In terms of the main requirements, the subject must obtain the work legally. In terms of purpose, it must be for non-commercial purposes. In terms of formal requirements, it is required to show respect to the copyright owner by indicating the source in the mining results. If there are orphan works, the database information of the source of the work should also be indicated [14]. While the United Kingdom stipulates that text and data mining can be used reasonably, it also stipulates transfer restrictions, purpose restrictions, and transaction restrictions to better protect the legitimate interests of copyright holders.

The copyright exception system for text and data mining in the UK, on the one hand, conforms to the basic content of the "three-step test method" embodied in the TRIPs agreement. Abuse also helps society rationally utilize the existing scientific and cultural achievements and realize the unification of personal interests and social interests.

3.3 European Union: relatively prudent exceptions

The European Parliament and the Council of the European Union approved the "Digital Single Market Copyright Directive" on March 26 and April 15, 2019, respectively. One of the goals of this directive is to adapt the exceptions

and limitations of copyright law to the development of the digital market. Among them, Articles 3 and 4 stipulate copyright exceptions to TDM, and the domestic laws of EU member states will establish TDM exceptions that meet the requirements of the "2019 Copyright Directive" within two years.

First of all, with regard to the applicable subjects, there are only two types of scientific research institutions and cultural heritage institutions. Then, regarding the applicable elements, the first is "legal source", that is, TDM users must have legal access to the mined text or data. Users who obtain texts or data illegally or illegitimately cannot benefit from the EU TDM exception [15]. The second is "safety protection measures". One of the applicable conditions of the EU TDM exception is that TDM users take appropriate security measures for the mined text and data. The lack of security protection measures will lead to the risk of copyright infringement of the mined content, and the TDM exception still applies at this time will result in the loss of rights holders' interests. The requirement of safe preservation measures can prevent the abuse of TDM exceptions and protect the legitimate interests of right holders [16].

4 The comparative law perspective of fair use of works in Machine Learning

4.1 The features of various countries' legislation

Generally speaking, the countries that stipulate copyright exceptions for TDM are basically developed countries in Europe and the United States, and generally have the following two characteristics: (1) Existing legislation basically limits TDM exceptions to the field of scientific research, (2) Text analysis and research activities that are exempt from liability under the TDM exception should be for non-commercial purposes. Among them, Japan is quite special. Its legislation does not explicitly limit the TDM exception to the use of works for non-commercial purposes, nor does it use the term "scientific research" with non-commercial implications.

A major development in the legislative practice of copyright exceptions for TDM is the Directive on Copyright in the Digital Single Market passed by the European Parliament in March 2019. Article 3 of the Directive stipulates a mandatory exception for TDM applicable to research institutions, which means that the copyright exceptions for TDM will be extended from countries such as Britain, France, and Germany to other EU countries, because EU members The country is obliged to transform the substantive content of the new EU Copyright Directive into domestic law within the statutory period. After the legal practice of many countries and the theoretical recognition of academic research, it can be considered that in non-commercial scientific research activities, TDM has gradually become a legal exception to copyright at the international level.

The United States is at the most lenient end in comparison, and its case law has basically agreed that the act of copying works by artificial intelligence does not constitute copyright infringement if there is no public

communication [17].

4.2 The enlightenment on the establishment of the fair use rule in China

The newly added "three-step test method" in Article 24 of the Chinese Copyright Law is semi-closed and is limited to "use in the following circumstances", which essentially constitutes a restriction on the interpretation of 12 specific situations, rather than an open authorization regulation. Courts cannot "interpret" new limitations or exceptions under this article in specific cases. In comparison, Article 107 of the U.S. Copyright Law adopts a flexible legislative model, which provides a broad and flexible space for interpretation by U.S. courts.

For text and data mining technology, the United States has a broad market and prospects, and has relatively mature judicial precedents [18]. The "conditional exception" represented by the UK and the EU is to limit some necessary content in the Fair Use system to better adapt to social development and respond to social voices [19].

So, there are two options for my country to optimize the legislative model of Fair Use: one is to adopt the flexible and open "Fair Use + specific enumeration" model of the United States, which stipulates four considerations for Fair Use. And then lists common Fair Use methods. The other is to follow the mainland. The "three-step test method + specific enumeration" model of the legal system clearly stipulates that "use for AI learning and creation" is Fair Use when amending China's "Copyright Law Implementation Regulations".

5 Conclusion

The Fair Use system has played an important role in coping with the development of information technology and maintaining the balance of interests many times throughout the history of copyright development. In the current situation which artificial intelligence has become a new engine of economic development and a new opportunity for social development, which is rapidly changing human life and becoming the focus of international competition. It will once again assume this important responsibility in China.

To sum up, this paper proposes adding a new type of Fair Use of machine learning after Article 24, Paragraph 1, Item 12 of China's current "Copyright Law", namely: "copying, adapting other people's works for machine learning, and using creation achievements are made available to the public in the form of broadcasting and information network dissemination, but the metadata of the obtained work shall not be publicly disseminated in its original form."

References

1. M. Sag, Journal of the Copyright Society of USA. The New Legal Landscape for Text Mining and Machine Learning. 66(2), 346–36 (2019)

2. L. Lessig, *Penguin Books*. Remix: Making Art and Commerce Thrive in the Hybrid Economy (2008)
3. H. Richter, Max Planck Institute for Innovation and Competition Research Paper. Position Statement of the Max Planck Institute for Innovation and Competition on the Mode of European Copyright Rules. (2), 7–10 (2017)
4. M. McJohn, L. McJohn, Northeastern University Law Review. Fair Use and Machine Learning. 12 (2020)
5. A. Li, Electronics Intellectual Property. Copyright Analysis of Machine Learning Works -- Non-Work Use, Fair Use, and Infringing Use. (06), 60–70 (2022)
6. Z. Xuan, Publishing Research. On Fair Use of Copyright in AI Creation from the Perspective of Classification Protection. (03), 81–87 (2022)
7. A. Foer, A. Durst, The Antitrust Bulletin. The Multiple Goals of Antitrust. 508 (2018)
8. HD. Wu, *A Study on the Fair Use System of Copyright*. China Renmin University Press (2020)
9. E. Lee, Southern California Law Review. Technology Fair Use. 83(4), 808 (2010)
10. H. Piwowar, J. Priem, V. Larivière, PeerJ. The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles. 6(4), e4375 (2018)
11. G. Spindler, International Review of Intellectual Property and Competition Law. Copyright Law and Artificial Intelligence. 50(9) (2019)
12. T. Pihlajarinne, International Review of Intellectual Property and Competition Law. Should We Bury the Concept of Reproduction—Towards Principle-Based Assessment in Copyright Law? 48(8) (2017)
13. Y. Bathae, Berkeley Technology Law Journal. Data privacy and user exception. 35(2020)
14. B. Sobel, Columbia Journal of Law & the Art. Artificial Intelligence Fair Use Crisis. 41(1), 49–79 (2017)
15. M. McJohn, L. McJohn, Northeastern University Law Review. Fair Use and Machine Learning. 12 (2020)
16. J. Boyle, *The Public Domain: Enclosing the Commons of the Mind*. Yale University Press. 29 (2008)
17. M. Lemley, 13Southwestern Journal of Law and Trade in the Americas. New Balance between IP and Antitrust. 2(3) 237-248 (2007)
18. D. Schönberger, Social Science Electronic Publishing. Deep Copyright: Up-and Downstream-Questions Related to Artificial Intelligence and Machine Learning. 10(1), 163-167 (2018)
19. C. Geiger, J. Griffiths, R. Hilty, European Intellectual Property Review. Toward a Balanced Interpretation of the "Three-Step Test" in Copyright Law. 30(12), 489–496 (2008)