WEP Warwick Evans Publishing

Transactions on Social Science, Education and Humanities Research
ISSN: 2960-1770, eISSN: 2960-2262 | Volume 7
ADCS 2024

Judicial Judgment Standard of Generative Artificial Intelligence and Training Data from the Perspective of Fair Use

Zhenyi Gao

School of Humanities and Social Sciences, University of Science and Technology Beijing, Beijing, China

U202143029@xs.ustb.edu.cn

Abstract. With the development of generative artificial intelligence, text and data mining has been put into more applications. However, as a technology for collecting and processing a large number of other works, it faces copyright infringement under the current legal system. As a step in the operation of artificial intelligence, text and data mining is to collect and process existing works and information to realize the behavior of knowledge "reproduction", thus promoting the utilization of knowledge and social innovation. Meanwhile, copyright exemption should be provided for its fair use system. The fair use system in the United States can give full play to the autonomy of judges in specific cases and flexibly identify fair use exemptions, while the copyright exception in Europe clarifies that text and data mining activities should be exempted under specific circumstances. Learning from both, China can not only break the closed fair use system, but also build an evaluation system of artificial intelligence infringement and benefit-sharing mechanism, so as to provide scientific norms for fair use and improve the development of artificial intelligence and big data industry in China.

Keywords: Text and Data Mining; Generative Artificial Intelligence; Fair Use.

1. Introduction

As for the application of generative artificial intelligence, it is necessary to analyze and collect relevant data within a certain range, identify and detect this part of data in real-time, and judge the internal relationship and data law to achieve the purpose of "learning". Such a technology is text and data mining (TDM). [1] In other words, the learning and development of generative artificial intelligence must be based on data mining.

As an emerging technology, generative artificial intelligence has a wide range of applications and development prospects, which also challenges the laws in China. [2] On the one hand, TDM encounters serious infringement risks in the context of China's copyright law. Besides, the development of the artificial intelligence industry is bound to require providing a legal basis for TDM. On the other hand, it is difficult for infringed copyright owners to safeguard their rights and interests through existing laws. In July 2023, the Office of Central Cyberspace Affairs Commission and other seven departments jointly issued the *Interim Measures for the Management of Generative Artificial Intelligence Services* for its standardization. In practice, there have been many disputes caused by the data mining of generative artificial intelligence, such as the case of Beijing Film Law Firm v. Beijing Baidu Netcom Science & Technology Co., Ltd. (a case of dispute over the infringement of protecting the integrity of works) [3], Shenzhen Tencent Co., Ltd. v. Shanghai Yingxun Technology Co., Ltd. (a case of dispute over the infringement of the copyright) [4], etc. Hence, TDM needs to be standardized urgently.

Given that text and data mining is facing serious infringement risks in the context of copyright law in China, fair use is a way to prevent the emergence of improper knowledge monopoly in China and is also one of the standards to judge whether a certain behavior infringes copyright, which is applicable to the norms of text and data mining behavior.

How to evaluate and deal with TDM is controversial in academia. Theoretically, there are disputes about whether TDM infringes a copyright and whether it should be exempted from copyright, the scope of the exemption and the judgment standard of fair use. In judicial practice, cases in which TDM infringes the rights of copyright owners or cases in which its generated results are protected by the copyright law all exist. Meanwhile, there are different patterns such as fair use and non-infringement exceptions to standardize TDM in foreign laws. This paper attempts to explain the legitimacy of TDM from the perspective of fair use and provides some suggestions for regulating this behavior in China.

2. Dilemma of TDM in China's Current Legal System

TDM as a new behavior confronts many difficulties in China. Theoretically, there are controversies on how to characterize TDM and judge whether it is illegal or needs to be exempted. In practice, lawsuits against generative artificial intelligence have appeared in China, but there is no legal norm to clearly define and judge the behavior of TDM. So current practice focuses on the application and judges according to the existing standards. However, in the current foreign laws, there are two patterns of providing copyright exemption for TDM including fair use and copyright exception.

2.1. Theoretical Disputes of TDM

TDM, as a new technical means, is still controversial in the current theoretical communities in China. There are many disputes over whether TDM involves copyright infringement, whether TDM needs to be provided with legal exemption, and what path to provide exemption for TDM.

TDM involves copying copyright owners' works, adapting others' works, and spreading them to the public. According to some scholars, these all violate the relevant provisions of the current copyright law, [5] while others hold that temporary copying does not belong to the scope of the right to copy in China, and TDM does not necessarily infringe the copyright law. [6] Furthermore, as for the influence of TDM, some scholars believe that generative artificial intelligence will have potential adverse effects on copyright owners and the market, which should be restricted through collective management and tax systems. [7] Meanwhile, other scholars also believe that TDM is necessary for scientific research and development, which should be exempted through fair use. [8]

This paper argues that fair use is the best way to provide exemption for TDM in China, but its realization and exemption scope are controversial. Based on the path of hermeneutics, some scholars realize its exemption through loose interpretation and expanded interpretation of existing legal provisions. [9] Others believe the legislative path is better, and TDM clauses should be added to the existing fair use exceptions to provide exemptions. [10] As far as the scope of the exemption is concerned, some hold that TDM should be restricted from the subject and purpose, which should be only for personal use or based on scientific research that constitutes fair use. [11] Moreover, some scholars believe that companies are an important force in the development of the generative artificial intelligence industry and commercial subjects should not be excluded.

In addition, as for the judgment standard of fair use, some scholars propose that whether data mining constitutes fair use should be determined by the "three-step test". [12] Some of the scholars hold that it is better to determine the four elements and transformative use. Meanwhile, the classification protection method is also recommended by some scholars for new construction of the composition elements and judgment standard of AI fair use. [13]

2.2. Insufficient Legal Norms of TDM

There are no special provisions on TDM in China's current legal system with a lack of comprehensive norms. According to the current legal system to guide the cases involving TDM behavior, it is difficult to avoid the situation that the behavior without infringement intention and infringement influence is convicted because it meets certain infringement requirements without committing a crime.

At present, China's laws and regulations targeting the regulation of generative artificial intelligence are the *Interim Measures for the Administration of Generative Artificial Intelligence* (hereinafter referred to as the *Interim Measures*), with only Article 7 involving TDM, while other laws do not mention generative artificial intelligence or TDM. It can be seen that the norms and evaluation of TDM in China can only be conducted by resorting to the general provisions of other laws.

It is also difficult to analyze TDM behavior according to the general provisions of existing laws. For example, Article 7 of the Interim Measures stipulates that "data and basic patterns with legal sources" should be used for training data activities, but it is not clear what sources meet this standard. During the data mining, it is necessary to copy data, which is suspected of infringing the copyright owner's right to copy. The process of data processing and the final output may infringe on the deductive right of the copyright owner. Transmitting data or text to the "cloud" or to the Internet during data mining may infringe the copyright owner's right to disseminate to the public. [14] However, due to the "black box" characteristics of generative artificial intelligence, copyright owners can only judge whether it is infringing according to its output results, and fail to know whether the artificial intelligence has implemented data mining on its works, nor can they identify the rights status of the mined objects.

There are three ways to resort to the fair use of works within the protection period by the public in China, including licensing, statutory licensing and fair use. However, it is obvious that a large number of generative artificial intelligence have not and cannot obtain text and data sources through licensing, [15] which does not conform to legal licensing. Although fair use is not completely listed, according to the provisions of Article 24, Paragraph 13 of the *Copyright Law*, whether it constitutes fair use still depends on the specific provisions of laws and regulations, rather than being left to the judicial discretion. The three-step test in the *Copyright Law* cannot provide an exemption for TDM either, because Article 24 of the *Copyright Law* already has a catch-all provision in Paragraph 13 about "other circumstances stipulated by laws and administrative regulations". Judging whether this judgment constitutes fair use is more like a method to assist the judiciary in judging, rather than as a supplement or a catch-all provision to Article 24. Moreover, the "three-step test" starts from the damage result and is fundamentally different from the provisions of Article 24 in terms of subject, purpose and form, which should be considered as a restriction on Article 24. In specific cases, the court cannot "interpret" new restrictions or exceptions according to Article 24.

2.3. Practice Controversy of TDM

With regard to cases involving generative artificial intelligence, China's judiciary mainly responds to lawsuits and determines whether it is infringement based on existing laws. In foreign laws, there are two main patterns including fair use and copyright exception.

At present, Chinese courts have adopted the traditional methods of identifying copyright infringement in generative artificial intelligence cases. For example, in the case of Tencent Co., Ltd. v. Shanghai Yingxun Technology Co., Ltd. (a case of dispute over the infringement of the copyright), [16] the judgement of the court was based on whether the work was original. In the case of Beijing Film Law Firm v. Beijing Baidu Netcom Science & Technology Co., Ltd. (a case of dispute over the infringement of protecting the integrity of works) [17], the court found that the processing of works by its artificial intelligence violated the plaintiff's right to protect the integrity of works. In these cases, the judge still evaluated the TDM behavior or generative artificial intelligence according to the infringement elements stipulated in the traditional copyright law, which has not considered the legal exemption to judge the case. In foreign laws, there are mainly two patterns to evaluate the TDM behavior and grant different degrees of exemption.

1. Fair Use Pattern

The United States provides certain exemptions for TDM through fair use. Whether it constitutes fair use is mainly judged by the four elements stipulated in *Copyright Law*, that is, by the purpose and nature of use, the nature of copyrighted works, the quantity and quality of copyrighted works occupied by use and the potential market or value influence of copyrighted works. In the case of Andy Warhol

Foundation for the Visual Arts, inc. v. Goldsmith et al. of Copyright Infringement, [18] the judgments and reasons made by the Court of First Instance, the Court of Appeal and the Supreme Court, which also judged Warhol through the four elements, are different. The decisive differences lie in the purpose and nature of trademark use as well as whether it constitutes transformative use. Moreover, the key lies in whether the new works encroach on the interests of the original works to a certain extent. The judgment of the four elements in the United States emphasizes "integrity" and solutions vary in different cases, so judges have a large room for interpretation and greater discretion with strong instability.

2. Copyright Exception Pattern

In 2019, the European Union adopted the *DSM Directive*, which adopted the copyright exception to regulate TDM behavior. This directive gives the exception of temporary replication and scientific research on TDM behavior. In other words, temporary and incidental exceptions without independent economic value are allowed and should be limited to "non-commercial purposes". In the Infopaq case, [19] the European Court of Justice held that the temporary copying involved in its data acquisition procedure constituted a copyright exception. It can be seen that the copyright exception of European Union has greatly restricted the subject and content of TDM, with its copyright exception only applicable to a small number of generative artificial intelligence. In contrast, the report *Impacts of Artificial Intelligence on Copyright and Patent System* released by the UK in 2022 allows data mining for any purpose (including non-commercial and commercial). At the same time, copyright owners still enjoy safeguard measures to protect their content. This report clarifies that the obligee can no longer charge for a TDM license in the UK but has the right to choose the platform to submit his works and provide his works to users through a subscription or a one-time fee.

3. Fair Use Pattern in China

China adopts a semi-closed doctrine in the form of fair use in legislation. In other words, the *Copyright Law* lists twelve specific situations with Article 13 as a catch-all provision which mentioned "other situations stipulated by laws and administrative regulations". However, there are no other specific provisions in China's laws and regulations, so only the aforementioned twelve situations constitute fair use at present. In the newly revised *Copyright Law* in 2020, there is a "three-step test" for fair use, which provides identification standards for the twelve listed situations, rather than providing identification standards for the fair use of all behaviors. Thus, only twelve situations can be applied to fair-use articles in China. Moreover, the "three-step test" is not crystal-clear and its guiding significance for practice is limited. Some scholars believed that China's fair use system is "closed" and "conservative". [20]

Although the legislation restricts fair use, there are different situations in practice. Due to the emergence of many new forms and technologies in practice, some behaviors that judges think should be exempted in practice are not listed in *Copyright Law*, so the court began to identify fair use according to the actual situation, which led to quite strong subjectivity. There are significant differences in the standards of fair use in various regions in China. More than two-fifths of the courts conduct flexible identification, mainly including two types of flexible identification standards. One standard is completely open and flexible, and the other is semi-closed. The former mainly identifies fair use through the "four elements" pattern of the United States, while the latter identifies fair use according to unspecified additional factors. This identification method has essentially broken through the existing legislative provisions. [21] The difference in judgments in various regions will damage the authority and credibility of laws.

In addition, due to the backwardness of legislation and lacking standards, fair use has been excessively applied in China's practice. The aforementioned scholars found that fair use was seriously abused in practice through empirical research on litigation cases in China. In more than 100 cases, only 9% of the fair use defense is established. Neither the court nor the public have a clear understanding of the determination of fair use. [22] Hence, there is an urgent need for the standard of fair use in China's judicature.

3. Justification of TDM Behavior and its Agreement with "Reasonable Interpretation" System

TDM behavior is about data collection and processing, which appears in the development of technology and is necessary for the development of industry. It does not include the purpose of damaging others' copyright, but has something in common with people's learning. Thus, it has the basis of legitimacy. Under the existing legal system of China, TDM behavior is the most suitable way to provide copyright exemption.

3.1. Justification of TDM Behavior

TDM behavior is the first step for generative artificial intelligence to realize "output", which aims to provide artificial intelligence with a data basis for analysis and "recreation". It is beneficial to the development of artificial intelligence and social innovation, which is consistent with the core purpose of the *Copyright Law of the People's Republic of China*. Meanwhile, TDM behavior meets the national development needs and the needs of people's well-being lives, which enables a legitimate basis for providing exemption.

TDM behavior includes data preprocessing, data analysis and mining and model evaluation. The data preprocessing of data mining is to clean, integrate, transform and reduce the original data, [23] which mainly aims to transform the mined text and data into a certain data format, and carry out preliminary screening and arrangement. Data analysis is to analyze the existing data in various ways to find rules. Model evaluation is to verify the level of generating a model for optimization. During these processes, they are mainly based on big data statistics and prediction-correction programs, which mainly include mature algorithms of five types of machine learning, including classification, regression, clustering, prediction and association, with algorithms varying in different generative artificial intelligence.

It can be seen that the direct purpose of TDM behavior is only to establish a database for artificial intelligence and find out the rules. TDM aims to process massive text and data, form its model, and then arrange and output results according to users' requirements. However, this process and its results are different due to various core algorithms. The algorithm condenses the labor value of artificial intelligence designers, so the algorithm plays a decisive role in the output results. Moreover, the text and data provided are used as materials to assist the algorithm in forming models. Hence, TDM is not for infringing the copyright of the obligee nor necessarily infringing the copyright.

From the perspective of the original intention of establishing Intellectual Property Law, as an auxiliary tool of society, it is to encourage more innovation, and infringement exemption is also established to urge innovation. The copyright system gives creators certain rights of "monopoly" to promote innovation. [24] As a tool of national social policy, TDM has rights to protect and the degree of protection are formulated according to the needs of national development. Meanwhile, public policies need to balance the rights and obligations of different subjects. When there are conflicts in many value objectives of copyright law, public interests have priority. [25] Absolutely speaking, copyright is not a natural right, and its establishment is just a compromise between the state and copyright owners to promote innovation. Some scholars even put forward "objectivism of works", advocating the protection of works rather than creators. [26] With a powerful transformative force, generative artificial intelligence even promoted the fourth industrial revolution and gave birth to new industries and new models, which is of great significance to the future development of China and mankind to a higher stage of civilization. [27] It is also a technology whose development is emphasized by China. According to the *Opinions of the Supreme People's Court on Giving Full Play to the Role of Intellectual Property Trial Function to Promote the Great Development and Prosperity of Socialist Culture and the Independent and Coordinated Development of Economy* issued by the Supreme People's Court, [28] TDM is favorable for innovative behaviors. China's State Council has also issued a document stressing a gap between China's artificial intelligence and developed countries, [29] thus the promotion of artificial intelligence is required. It is necessary to provide inclusive and flexible legislative orientation for TDM behavior, [30] which is inclined by intellectual property law.

From a technical perspective, artificial intelligence needs to be inclusive in data mining. No matter how mature the model is, it is impossible to form a high-level generative artificial intelligence under the current situation without massive data for training. Taking the well-known ChatGPT as an example, it uses hundreds of billions of tagged data for training. The amount of training data will affect the richness and perfection of "thinking" of generative artificial intelligence, which will affect its final application effect. As for generative artificial intelligence, the "marginal increasing" effect of data volume shows the importance of data volume for developing artificial intelligence. At present, when artificial intelligence is rising in various industries, it is increasingly becoming a vital production tool, giving copyright exemption to data mining behavior with data inclusiveness for the development of the artificial intelligence industry.

Some scholars believe that the data mining of artificial intelligence is not completely justified. As a profit-seeking product, artificial intelligence itself will infringe on the rights of copyright owners with potential adverse effects on the market. Copyright owners should be protected and large-scale artificial intelligence technology companies should be restricted through collective management and tax systems. [31] However, the above viewpoints are difficult to stand on either the theoretical basis or the solutions. First of all, artificial intelligence does not necessarily lead to infringement of copyright owners. The original intention of generative artificial intelligence is to create content, which is fundamentally different from the old artificial intelligence. It is overkill to limit its data mining behavior only because of the possible infringement danger of "output". Secondly, the purpose of setting up intellectual property rights is to stimulate the production of intellectual products, which is not justified to allow copyright owners to monopolize their copyrights for data mining. [32] Thirdly, as far as collective management is concerned, when artificial intelligence needs massive data mining, it is impossible to require it to obtain all permissions in advance, and the lack of data is not conducive to the development of artificial intelligence.

From the perspective of social public interests, some scholars hold that the too loose copyright protection pattern will put creators at a disadvantage and make it difficult to defend their rights, which is not conducive to producers' creation, thus impairing social public interests. [33] Some scholars also believe that data mining is conducive to the development of information collection technology and the promotion of public interests. [34] If China restricts data mining behavior too much, it may lead to a "chilling effect" in technology, and new technologies flow to more environmentally friendly countries, which hinders China's participation in international competition. [35] Some regard it as a dilemma. If the development of artificial intelligence is not restricted, it may undermine the creative enthusiasm of human beings. [36] In fact, from a longer-term perspective, generative artificial intelligence as a tool can be used to assist humans in creating. From a macro perspective, this technology promotes the dissemination, utilization and reproduction of knowledge. In the era of artificial intelligence, TDM has a wide range of use space, [37] which can produce new knowledge and new value. [38] In the post-information age, a large amount of data is generated and transferred at all times. Meanwhile, data is already a direct wealth and social resource. With the application of data technology, new knowledge can be discovered, new value can be created, and the leap from data to knowledge and from knowledge to action can be realized. The demand for data utilization in public and private fields is more urgent than in any previous era. Generative artificial intelligence is one of the most vital and efficient ways to collect this information and create value. If text and data mining are not exempted from fair use, the money and time paid by artificial intelligence for obtaining copyright will be greatly increased. The lack of data will seriously affect the effectiveness of artificial intelligence, which hinders the efficiency improvement of knowledge utilization. Supporting TDM is beneficial to promoting artificial intelligence and value creation. In addition, Article 5 of the *Interim Measures* also reveals the tendency to encourage the development of generative artificial intelligence. [39]

3.2. Discussion on the Agreement of TDM and "Fair Use" System

Throughout the system of copyright exemption for legal acts under the current legal system in China, TDM is the most suitable for fair use. On the one hand, generative artificial intelligence utilizes existing works through new technical means and innovates the way of using works, which can not only use existing works, but also promote the public understanding and knowledge of works. All these are conducive to promoting the dissemination and development of knowledge, which is consistent with the purpose of fair use. In addition, as for the form of infringement, it is similar to the fair use system. Meanwhile, how to bring it into China's fair use system needs to be discussed.

As a system in the intellectual property law, the purpose and significance of establishing a fair use system lies in preventing the unfair knowledge monopoly behavior, protecting the fairness of public learning and accumulating knowledge, and improving the efficiency of knowledge utilization within a reasonable range by restricting copyright. The fair use system is a tool to balance the interests of creators, disseminators and users, [40] which is also a tool to balance personal interests and public interests. From a limited perspective, what acts are granted with copyright exemption and how much exemption is granted are the trade-offs between legislators and judiciaries on the interests of copyright owners and the public welfare in obtaining and using knowledge. From the perspective of purpose, TDM is the early behavior of generative artificial intelligence. Through the accumulation and processing of data, the model is trained, which provides the basis for generating the new content of artificial intelligence. The purpose of this behavior is to assist the "output" of generative artificial intelligence, that is, to promote new content review. Through the computer to deal with text and data, we can find out the connection in massive data at an extremely fast speed, and provide users with a large amount of information. Giving fair use exemption to TDM can greatly save a lot of money and time paid by artificial intelligence to obtain copyright, which helps improve the high efficiency of knowledge utilization and conforms to the purpose of a fair use system.

From the formal perspective, China's current fair use system judges whether to grant exemption according to Article 24 and Article 21 of the *Copyright Law* under the condition of infringing others' copyright. TDM behavior is a part of the operation of generative artificial intelligence, which is no longer private until the "output" result is produced. Copyright owners also claim their rights according to the output result, which is similar to the general copyright infringement claim. In judicial practice, there have been cases where the output results of generative artificial intelligence are regarded as the works of natural persons to consider the case, such as Shenzhen Tencent Co., Ltd. v. Shanghai Yingxun Technology Co., Ltd. (a case of dispute over the infringement of the copyright). [41] The contradiction of TDM in copyright law is the conflict between copyright owners and artificial intelligence users and creators. The three-step test for judging whether it constitutes fair use in China's legislation stipulates that the spirit of "not affecting the normal use of the work, but also reasonably damaging the legitimate interests of the copyright owner" can be used to guide and coordinate this contradiction. There is no obvious incompatibility between TDM and the current fair use system.

Although TDM and the fair use system have a high degree of agreement, it is not appropriate to directly bring TDM into the current fair use system in China without amending the legislation. Some scholars tried to explain the data mining behavior from the existing reasonable interpretation provisions in China, but they inevitably resort to the expanded interpretation of legal provisions and the split of data mining behavior. For example, Lin Hua believed that "input" and "output" were independent of each other, which could be applied separately to create a context for data input that only involves the right of reproduction, and apply "individual" with reference to the "organization" to provide a legal explanation for data mining. By distinguishing for-profit subjects from non-profit ones, Wan Yong made a loose interpretation of "a small amount" and an expanded interpretation of "scientific research", so as to provide a reasonable interpretation of data mining behavior. However, the above-mentioned method, whether it is to distinguish subjects or decompose behaviors, is to split the behavior of data mining with the same nature and embed it into the existing regulations. TDM behavior is carried out by machines for the "creation" of generative artificial intelligence, which requires a large amount of data processing, and is often conducted by non-organizations or institutions

with a high probability of commercial purposes. [42] It is apparently different from Article 1 and Article 16 of the *Copyright Law*, which are for personal study or scientific research respectively. In addition, this interpretation can only provide a legal basis for the existing artificial intelligence data mining behavior based on the current understanding of artificial intelligence, which is out of touch with its future development. Wan Yong also believed that this was only a "measure for emergency". As for for-profit artificial intelligence data mining and machine learning, "a better way to provide a normative basis is still to amend legislation." [43] As a new behavior of applying works, TDM has a large number of applications in the future and even now. It is quite different from the existing fair use situation in terms of purpose and nature, which is qualified and should be regarded as a separate fair use situation.

3.3. Further Discussion on the Exemption System of TDM in China

Since it is determined to provide copyright exemption for TDM through fair use, other problems need to be solved. This paper argues that when a fair use system is applied to TDM, the design of this behavior is dominated by artificial intelligence companies and manufacturers, so its exemption subjects should be both. As for two patterns in foreign laws, fair use is more conducive to providing exemptions for specific cases flexibly and openly, while the copyright exception will inevitably trigger a more closed and conservative judgment because of its position and involving methods. This paper holds that a fair use pattern is better to promote innovation in China and strive for greater advantages in new fields.

To provide fair use exemption for TDM, the exemption subject should be identified first. Artificial intelligence does not have self-awareness, self-behavior control and resources now, which does not conform to the constituent elements of natural science subjectivity. From the perspective of legal value, artificial intelligence is not only unnecessary to be a legal subject, but should be regulated from the perspective of preventing its risks to humans. Without independent consciousness, property and personal foundation, artificial intelligence fails to conduct behavior and shoulder the responsibility, thus not constituting a legal subject. As an ability, artificial intelligence belongs to the extension of its subject's behavior or the way of behavior, which also realizes the value of its maker's or user's behavior, so it belongs to the elements of its subject's legal behavior in law. [44] Artificial intelligence does not belong to and should not constitute a legal subject in the current legal system. Generative artificial intelligence involves various natural persons, including digital copyright owners, publishing institutions, technology providers and consumers. [45] This paper argues that the applicable subject should be companies rather than individuals. Besides, the obligation to prevent artificial intelligence from generating output results improperly infringing others' copyrights lies with companies that have technological advantages and provide such a service. Some scholars have mentioned the feasibility of algorithms' fair use. [46] Undoubtedly, this does not affect the necessity of exempting and regulating this behavior by amending the law. No matter what the technology is, the law needs to regulate it from a more realistic macro perspective through various considerations, and the algorithm can achieve "better" performance rather than being "perfect". Moreover, there is a main problem that artificial intelligence companies need to consider the design of this algorithm needs to be based on legal norms.

To promote innovation, AI R&D personnel should not only be allowed to fully use existing materials, edit and create more new databases, but also improve the AI technology and reduce the transaction cost of using copyrighted materials. There are two legislative patterns for optimizing fair use in China. Firstly, we should learn from the flexible and open pattern of "fair use + specific enumeration" in the United States, stipulate four elements of fair use, and then list common fair use methods. Secondly, following the pattern of "three-step test + specific enumeration" in the civil law system. It is clearly stipulated that "the use of AI learning and creation" belongs to fair use when revising the *Regulations for the Implementation of Copyright Law in China*.

In terms of the legal system, the fair use pattern in the United States mainly follows the existing judgment standard of "four elements" in judicial practice without legislative amendment. Facing the

new use of other people's works in practice, there is no need to legislate, which has little impact on the original legal system and is conducive to maintaining the seriousness and authority of the law. In the application of law, this pattern can apply the law flexibly and is not constrained by the provisions. In the case of Google LLC v. Oracle America, Inc. (a case of dispute over copyright infringement), [47] the Supreme Court of the United States proposed that the judgment of fair use of this case was based on the assumption that Java API was protected by copyright law. In other words, it bypassed whether Java API was protected by copyright law and directly judged through "four elements". In judicial practice, this pattern pays attention to considering the whole and emphasizes the balance of interests. In the case of Campbell v. Acuff-Rose Music, Inc., [48] the court first denied that "use for commercial purposes" must be "unreasonable", and the purpose is only one of the measurement standards of "four elements". This pattern can be result-oriented and provide more autonomy for various innovations. However, it lacks clear standards for subjects using other people's works simultaneously, especially through new ways. For example, as for TDM, users can't accurately know whether their use constitutes fair use. In addition, this pattern gives the court great discretion, which easily leads to the risk of expanding interpretation and abusing fair use.

The European Union's pattern of copyright exception stipulates the fair use situation in data mining through legislation. AI companies have clear criteria for judging, which is known that temporary reproduction of works for non-commercial purposes can constitute fair use. As for judges, this pattern has an accurate basis for judgment, which can be targeted and accurately applied to the subject of fair use in judicial adjudication, making the adjudication results more unified. As for other subjects in the market, it plays a guiding role, and copyright owners have clear standards for under what circumstances they can safeguard their rights. However, this legislative pattern lacks flexibility, and artificial intelligence for TDM is a new behavior, which may undergo new changes in its form, use and mechanism with limitations in technological and social development. In addition, its extensibility is questioned. [49] European Union and China belong to the continental law system. Fair use in China is regulated by legislation now without regulation on TDM behavior.

Table 1. Comparative Table of TDM Standard System Between the United States and the European Union

	United States	European Union
Judgment Standard	Four Elements	Three-Step Test
Judgement Scale	Loose	Strict
Judge's Discretion	Larger	Smaller
Scope of Exemption	Open	Closed

According to this table, the fair use system in the United States is more like exempting all substantive and legitimate behaviors from the user's perspective, thus promoting innovation. The European Union, on the other hand, takes a prudent attitude towards copyright from the perspective of the obligee and does not constitute crimes except those acts that have been identified as should be exempted, so as to protect the rights and interests of the obligee. The European Union's position stems from its high protection of personal information, which recognizes both personal information and privacy as basic rights and creates personal information autonomy. According to the case of Google Inc. v. Agencia Espanola de Proteccion de Datos (AEPD) & Costeja Gonzalez, [50] the European Court of Justice puts personal information and privacy ahead of the company's economic interests and the public right to be informed when balancing interests. Relatively speaking, the loose scope of fair use in Britain and the United States is influenced by mercantilism, and they understand privacy from the perspective of freedom. In the Electronic Privacy Information Center v. Facebook, [51] the judge judges whether the company's activities constitute an infringement exception by analyzing whether they constitute

"daily business activities". Hence, privacy rights do not take precedence over business development in the United States, but need to weigh two interests in specific cases. Fair use could cover "full reproduction" in The Authors Guild, Inc. v Google, Inc. [52] This flexible regulation provides institutional autonomy for the new and revolutionary use needed by the development of artificial intelligence. The vigorous development of large databases and artificial intelligence in the United States is intertwined with its good institutional environment.

At present, China has not yet formed a systematic and concrete personal information protection system, which just reduces the shackles when balancing the interests of personal information and industrial development. In current society, information contains great commercial value and public management value. Besides, the legitimacy of collecting and utilizing personal information by information providers has been recognized by legislation and society. Text and data mining has an urgent need for information from the perspective of generative artificial intelligence, so it should be moderately loosened when constructing the scope of its fair use system, which is in line with the requirements of developing China and the essence of intellectual property law.

There are two main factors for the adoption of the fair use pattern in the United States. First of all, its legal system is a case law system, and the adjudication process of judicial organs is also a legislative process, which can maintain the unity of law through adjudication standards. Secondly, the legislative process in the United States is complicated. After a new bill is put forward, it needs to be reviewed by professional committees, discussed by the House of Representatives, reviewed by the Senate and House of Representatives respectively, and then negotiated and revised to form a unified text, which can be signed by the president before a law can be formed. Meanwhile, most of the bills can't even pass the Committee's consideration. This also involves the reconciliation of the Senate and the House of Representatives as well as the reconciliation of different parties, so it needs a long cycle and repeated review and revision, and it is difficult to legislate. Its most prominent advantage is that it can flexibly face the changes in practice and avoid the obstacles that need to be bypassed in legal practice, which is insufficient for fair use in China at present. TDM by generative artificial intelligence is a new and developing technology, with its specific methods and contents likely to change. A relatively flexible and result-oriented mechanism can better meet the practical needs and future development of this technology and justice. At present, China's fair use system has some shortcomings in flexibility and openness. Learning from the fair use pattern of the United States can provide support for judicial practice and industrial development. In addition, as for the legislative pattern, China does not have such complicated procedures as the United States and mutual constraints between the two parties and the two houses, so legislation is relatively easy. Besides, there has been enumerated fair use in the legislation. If TDM into the specific circumstances of fair use will help to overcome the lack of accuracy of the fair use system in the United States. The introduction of this system can also change the closed defects under the current system, which is conducive to coping with future changes.

4. Legal Support Construction and Evaluation System Reconstruction of Data Mining Behavior from Four Aspects

This paper argues that the norms of TDM behavior can be improved from four aspects. First of all, to maintain the stability of the existing legal structure and clarify the legitimacy of TDM behavior, a specific exception of fair use should be added to Article 24 of the Copyright Law. Secondly, to break through the closed legislative pattern at present and embrace future changes for promoting national innovation and development, we should refer to the fair use standard of the United States to formulate norms for determining fair use and grant judges the discretion to exemptions via this standard. Finally, based on the public welfare of society, the benefit-sharing mechanism of artificial intelligence is established through extra taxation.

4.1. Add Specific Exceptions to Fair Use

In China's current legal system, data mining behavior lacks legitimacy support. As the most basic link of generative artificial intelligence, data mining behavior should be affirmed as soon as possible. Based on the previous discussion, data mining behavior is most suitable to be characterized by a "fair use" system, and the incomplete enumeration legislative form in Article 24 of China's *Copyright Law* gives data mining behavior a possibility to add to the existing fair use system. Although Paragraphs 1 and 6 of the existing specific exceptions have a certain correlation with data mining behavior, it aims to serve generative artificial intelligence with different purposes but the same data mining behavior. Moreover, data mining is quite different from the original paragraph in purpose and nature. To include data mining activities in these two Paragraphs, an extended interpretation of their original concepts would be involved, so it would not be appropriate to be added to the existing exceptions.

Data mining behavior is one of the most typical and special behaviors in the era of artificial intelligence, which plays a fundamental and decisive role in generative artificial intelligence. It is a behavior that will inevitably exist in large quantities in the future, which can be regulated by a separate regulation. In addition, data mining is a new technology and it is difficult to predict what new behaviors will be brought about by the development of artificial intelligence. A separate paragraph on data mining behavior in the *Copyright Law* is also conducive to making appropriate adjustments without destroying the formed system in its development. To sum up, the author believes Paragraph 13 can be added to Article 24 of the *Copyright Law*, which stipulates that "the data mining conducted by generative artificial intelligence has not caused unreasonable damage to the copyright owner." The original Paragraph 13 is postponed to Paragraph 14.

4.2. Break the Closed Legislative Pattern of Fair Use

A contradiction exists between the closed fair use pattern and the flexible and fair use in judicial practice. [53] The current TDM has revealed that the practice is ever-changing and the closed legislative pattern in essence will limit the development of technology in practice in the rapid development of technology at this stage. [54] However, copyright, as a right established to encourage innovation, may deviate from its original intention in essence by protecting it with a too-conservative attitude. Thus, new standards should be set up in the law, and fair use exemption can be obtained through behaviors that meet the judgment standards. Nevertheless, several paragraphs in Article 24 of the *Copyright Law* are only specific enumerations applying fair use.

4.3. Establish an Evaluation System for Artificial Intelligence Infringement

Generative artificial intelligence may infringe others' copyright, which is mainly related to its model. The direct purpose of data mining is to form a database for generative artificial intelligence and process data. No matter whether the output result of the generative artificial intelligence is infringing or not, the only direct purpose of data mining behavior itself is to generate a database. It seems that it is of little significance to analyze the specific methods of data mining separately and concretely.

Furthermore, the evaluation of data mining behavior cannot be separated from the generative artificial intelligence it serves. Although the data mining behavior is no different in nature, the generative artificial intelligence may infringe. For those generative artificial intelligence outputting infringing content due to poor model, the punishment should not be limited to the output, but resort to the input. Meanwhile, its data mining behavior should be restricted or prohibited simultaneously. Otherwise, the infringement of generative artificial intelligence can only be remedied afterwards. Some scholars proposed that characterizing TDM behavior and artificial intelligence generation behavior separately will make the subsequent legitimacy determination meaningless and trigger the deviation of fact restoration and logical fault. [55]

Therefore, whether the data mining behavior is infringing is consistent with the determination of whether the output of the generative artificial intelligence is infringing. As to whether the data mining behavior is infringing, we should mainly start by investigating the damage results and the similarity

between the output results and the original works. Based on the existing "three-step test", we can draw lessons from the "four elements" for analysis to determine whether the data mining behavior is infringing.

In addition, from the perspective of information sources of data mining, its sources cannot be obtained by illegal means. For example, if the information of data mining involves the privacy of others, it will infringe on their privacy. From the aspect of the data mining process, the owner of artificial intelligence has the obligation to do a good job in protecting the process of data mining and avoid the information leakage of copyright owners.

4.4. Introduce a Benefit Sharing Mechanism

Generative artificial intelligence is a model created by human beings by imitating their thinking, but human thinking is not completely transparent and we do not fully understand human thinking. After all, artificial intelligence is not human beings. Whether its output results can be called "works" and fully conform to "creativity" is still controversial. Some scholars believe that the current artificial intelligence is still in the stage of "weak artificial intelligence" as an intelligent system with unilateral capability. [56] In any case, the current generative artificial intelligence, as a service or commodity, is still a profitable tool for producers. Producers make profits with the knowledge achievements of the whole society as the means of production. Without social knowledge achievements, artificial intelligence is difficult to develop. They have the responsibility to give back to society, and enterprises themselves have the obligation to fulfill their social responsibilities. On the premise that it is impossible to pay for all copyright owners, taxation is essentially the best way. Considering the close relationship between generative artificial intelligence and knowledge, the author advises to establish a benefit-sharing mechanism and collect certain property from artificial intelligence enterprises by taxation, which can be used to encourage social knowledge creation.

5. Conclusion

Text data mining technology supported by computer software and big data has become the basic tool for intelligent development of all walks of life in the digital age, which has also been paid attention to by the Communist Party of China and the state. However, this behavior is faced with difficulties such as suspected infringement and the lack of an infringement evaluation system under the current legal system in China. In this paper, the most appropriate measure in China at present is to bring data mining behavior into the "fair use" system, affirm data mining behavior by adding specific exceptions in Copyright Law, establish an evaluation system of artificial intelligence infringement in judicial practice, and introduce a benefit sharing mechanism to realize the support and management of data mining. How to determine the amount of specific benefit sharing, and whether the results of data mining and the "output" results of generative artificial intelligence have copyright need further discussion.

References

[1] Jiang, Y. P. (2023). Application of data mining technology in network security. China CIO News, (05): 73-75.

[2] Zhang, H. B. & Xiao, Q. X. (2021). The construction of copyright exemption rules for text and data mining in the era of artificial intelligence. Science Technology and Law (Chinese-English Version), (06): 74-84.

[3] Beijing Internet Court. Civil Judgement. (2018) Jing 0491 Min Chu 239.

[4] Civil Judgment of People's Court in Nanshan District, Shenzhen City, Guangdong Province. (2019) Yue 0305 Min Chu 14010.

[5] Wu, H. D. (2020). Fair Use System of Copyright. Beijing: China Renmin University Press, 235-236, 312, 290.

[6] Zhu, L. (2007). Whether temporary reproduction belongs to reproduction in the sense of copyright law—Normative analysis with international conventions as the core. Electronics Intellectual Property, (01): 4.

[7] Hua, J. (2019). The dilemma and solution concerning application of copyright exceptions to artificial intelligence's creation. Electronics Intellectual Property, (04): 29-39.

[8] Zhang, H. B. & Xiao, Q. X. (2021). The construction of copyright exemption rules for text and data mining in the era of artificial intelligence. Science Technology and Law Chinese-English Version, (06): 74-84.

[9] Wan, Y. & Li, Y. L. (2023). Research on the interpretation of fair use clause in response to the development of artificial intelligence industry. Digital Law, (03): 83-92.

[10] Liu, S., Ren, J. D., Pi, Z. X. & Chen, Z. Y. (2023). Research on data mining technology based on AI technology under digital background—Comment on the principle and application of artificial intelligence and data mining. Chinese Sciencepaper, 18(06): 704.

[11] Wu, G. & Huang, X. B. (2021). Study on the design of fair use rules for text and data mining in the age of artificial intelligence. Library and Information Service, 65(22): 3-13.

[12] Yang, X. D. (2020). Research on the fair use of artificial intelligence editing. Science Technology and Law (Chinese-English Version), (01): 8-14.

[13] Xuan, Z. (2021). On the fair use of copyright in artificial intelligence creation from the perspective of classified protection. Publishing Research, (04): 81-87.

[14] Wan, Y. (2021). Dilemma and solution of fair use system of copyright law in the era of artificial intelligence. Social Science Journal, (05): 93-102.

[15] Wan, Y. & Li, Y. L. (2023). Research on the interpretation of fair use clause in response to the development of artificial intelligence industry. Digital Law, (03): 83-92.

[16] Beijing Internet Court. Civil Judgement. (2018) Jing 0491 Min Chu 239.

[17] Civil Judgment of People's Court in Nanshan District, Shenzhen City, Guangdong Province. (2019) Yue 0305 Min Chu 14010.

[18] Andy Warhol Foundation for the Visual Arts, inc. v. Goldsmith et al.

[19] C-5/08 Infopaq International A/S v. Danske Dagbaldes Forening.

[20] Qin, J. (2022). Research on internet fair use system under the background of innovation driven—Also on the revision proposal of Article 24 of the copyright law. Science Technology and Law (Chinese-English Version), (05): 76-84+122.

[21] Diligena, D. & Song, X. T. (2022). The application of fair use in the judicial practice of copyright in China—Empirical analysis based on 113 litigation cases. Science Technology and Law (Chinese-English Version), (01): 109-117.

[22] Beijing Internet Court. (2018). Civil Judgement. 0491(239).

[23] Liu, S., Ren, J. D., Pi, Z. X. & Chen, Z. Y. (2023). Research on data mining technology based on AI technology under digital background—Comment on the principle and application of artificial intelligence and data mining. Chinese Sciencepaper, 18(06): 704.

[24] Wang, Q. (2021). A Course in Intellectual Property Law (Seventh Edition). Beijing: China Renmin University Press, 284.

[25] Lin, X. Q. (2021. Reshaping the fair use system in copyright law in the AI era. Chinese Journal of Law, 43(06), 170-185.

[26] Yi, J. M. (2017). Is creation generated by artificial intelligence work? Science of Law (Journal of Northwest University of Political Science and Law), 35(05): 137-147.

[27] Dong, K. Y. & Zhang, Y. Q. (2023). Philosophical reflection on the development and governance of generative artificial intelligence. Journal of Fujian Normal University (Philosophy and Social Science Edition), (04): 48-63.

[28] "Under special circumstances necessary to promote technological innovation and commercial development, considering factors such as the nature and purpose of the use behavior of the work, the nature of the used work, the quantity and quality of the used part, and the impact of the use on the potential market or value of the work, if the use behavior neither conflicts with the normal use of the work nor unreasonably damages the legitimate interests of the author, it can be regarded as fair use."

[29] Notice of the State Council on Printing and Distributing the Development Plan of the New Generation. [2017]35.

[30] Wu, H. D. (2020). Rethinking the copyright of works generated by artificial intelligence. Peking University Law Journal, 32(03): 653-673.

[31] Hua, J. (2019). The dilemma and solution concerning application of copyright exceptions to artificial intelligence's creation. Electronics Intellectual Property, (04): 29-39.

[32] Wang, K. W. (2021). Artificial intelligence data input and fair use of copyright. Journal of Library and Data, 3(02): 110-118.

[33] Liu, Y. H. & Wei, Y. S. (2019). The copyright infringement issue of machine learning and its solutions. ECUPL Journal, 22(02): 68-79.

[34] See Krista Cox. (2015). Fair use in text and data mining: ARL publishes issue brief. Association of Research Libraries. Retrieved from https://www.arl.org/news/fair-use-in-text-and- data-mining-arl-publishes-issue-brief/.

[35] Wang, W. M. (2022). Challenge and response of artificial intelligence for the restriction and exception rules of copyrights. Journal of Law Application, (11): 152-162.

[36] Sobe, L. B. (2018). Artificial intelligence's fair use crisis. Columbia Journal of Law & The Arts, (1): 45-97.

[37] Zhou, L. L. (2017). Study on European Commission's copyright exception for researchers and text and data mining. Library Development, (07): 19-24+30.

[38] Zhang, Z. Y. & Zhou, L. (2015). The concept, method and development path of big data publishing. Publishing Research, (01): 14-17.

[39] "Encourage the innovative application of generative artificial intelligence technology in various industries and fields … Support industry organizations, enterprises, educational and scientific research institutions, public cultural institutions and relevant professional institutions to cooperate in the innovation of generative artificial intelligence technology, data resource construction, transformation and application, risk prevention, etc."

[40] Gao, L. (2023). Inspection and reconstruction of copyright reasonable use system in the digital age: Theoretical analysis based on technology neutrality. Journal of Soochow University (Law Edition), 10(01): 41-52.

[41] Civil Judgment of People's Court in Nanshan District, Shenzhen City, Guangdong Province. (2019) Yue 0305 Min Chu 14010.

[42] Tang, S. H. (2017). Research on copyright exceptions of text and data mining in data environment from the perspective of EU DSM copyright directive proposal. Intellectual Property, (10): 109-116.

[43] Wan, Y. & Li, Y. L. (2023). Research on the interpretation of fair use clause in response to the development of artificial intelligence industry. Digital Law, (03): 83-92.

[44] Cao, W. (2023). Humanistic logic and legal analysis of artificial intelligence from the perspective of technological iteration. Renmin University Law Review, (01): 104-124.

[45] Li, D. S. (2021). Legal Construction of Digital Publishing in China from the Perspective of Intellectual Property Protection. Intellectual Property Publishing House.

[46] Shao, H. H. (2023). Copyright algorithmic fair use: Necessity, possibility and limitation. Law and Economy, (04): 149-164.

[47] Google LLC v. Oracle America, Inc. 593 U.S.

[48] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569.

[49] Rosati, E. (2019). Copyright as an obstacle or an enabler? A European perspective on text and data mining and its role in the development of AI creativity. Asia Pacific Law Review, (2):198-217.

[50] Google Spain SL & Google Inc v. Agencia Espanola de Proteccion de Datos (AEPD) & Costeja Gonzalez, C-131/12.

[51] Electronic Privacy Information Center v. Facebook., 3:15-cv-03747.

[52] The Authors Guild, Inc. v Google, Inc., 804 F. 3d 202, 2nd Cir. (N. Y.), Oct. 16, 2013.

[53] Xiong, Q. (2019). Transformative use interpretation in China copyright law. The Jurist, (02): 124-134+195.

[54] Google LLC v. Oracle America, Inc. 593 U.S.

[55] Jiang, K. (2015). "Conduct" in fair use stage. Law Review, 33(06): 185-193.

[56] Li, S. & Yang, Q. (2023). Can Turing machines pass Turing test? --From the perspective of the material theory of induction. Chinese Journal of System Science, (02): 28-32.