WEP Warwick Evans Publishing

Transactions on Social Science, Education and Humanities Research
ISSN: 2960-1770, eISSN: 2960-2262 | Volume 14
LHPPR 2025

# Research on the Adjudication of Infringement in the Use of Works in AI Data Training and the Statutory License System

Xucan Liu *

School of Intellectual Property, East China University of Political Science and Law, Shanghai, China

* Corresponding Author Email: a15011296135@163.com

**Abstract.** In response to the national policy directive on improving the development and governance mechanisms for generative artificial intelligence, and to address the challenges posed by AI technology to the current copyright system, this paper focuses on the core legal issues surrounding the use of works in AI model data training. The research first deeply analyzes the nature of the use of works during the data collection and model training stages, demonstrating that constitutes acts of use regulated under the current copyright law. Thus, unauthorized use constitutes an infringement upon the exclusive rights of copyright holders, such as the right of reproduction. This paper, grounded in an analysis of the operational mechanisms of data training in artificial intelligence, argues that the use of copyrighted works during the stages of data collection and model training constitutes a form of use subject to regulation under existing copyright law. Accordingly, the unauthorized use of such works by AI technology companies for training purposes, absent the consent of the rights holders, shall be deemed an act of infringement. This paper contends that neither the traditional licensing model nor the proposed fair use approach can adequately resolve the tension between the protection of copyright holders' rights and the advancement of artificial intelligence technologies. An overly stringent intellectual property regime may undermine the interests of rights holders, while an excessively permissive framework risks stifling technological innovation. To address this dilemma, the paper advocates for the appliance of statutory license to the field of data training. Such a regime would enable AI technology companies to make large-scale use of copyrighted works while ensuring that the legitimate interests of rights holders are duly respected. Besides, this approach aligns with the dual value orientation of copyright law, which seeks to balance protection with limitation.

**Keywords:** Artificial Intelligence; Data Training; Copyright Infringement; Statutory License; Balance of Interests.

## 1. Introduction

In recent years, artificial intelligence (AI) technology, represented by deep learning, has developed rapidly, with one of its core driving forces being large-scale data training. In this process, to enhance model performance and intelligence levels, developers often need to use vast quantities of existing works, such as texts, images, and audio-visual materials, as training data. However, while this use of others' works promotes technological advancement, it also inevitably touches upon traditional copyright law boundaries, sparking widespread discussion and intense focus on the core legal issue of whether the use of works in AI model data training constitutes infringement and how it should be regulated.

Properly resolving the issue of using works in AI data training is of crucial practical significance for balancing the legitimate rights and interests of copyright holders with the innovative development of the AI industry. Overly strict restrictions on the use of works may impede the iteration and application of AI technologies; conversely, excessively permissive unauthorized use will severely erode the incentive mechanisms for creators, damage the foundations of cultural creation, and ultimately prove detrimental to the long-term healthy development of technology. Currently, academia and judicial practice have not yet reached a consensus on key issues such as the legal characterization of the use of works in data training, infringement determination standards, and liability attribution. Existing

theoretical paths, such as "non-work-related use," "fair use," and traditional "authorized licensing," have all shown certain limitations and applicability difficulties when attempting to respond to this challenge, making it difficult to effectively bridge the tension between copyright protection and technological development.

This paper, grounded in the fundamental principles of copyright law and the practical characteristics of AI data training, adheres to the dual value orientation of "equal emphasis on protection and limitation." Firstly, this paper will deeply analyze the specific modalities of using works during the data collection and model training stages, demonstrating that under the current copyright law framework, unauthorized use constitutes an infringement upon the exclusive rights of copyright holders, such as the right of reproduction. Secondly, based on a critique of the inadequacies of existing theories, this paper advocates for the introduction of a statutory license system in the field of data training as the preferred path to balance various interests and resolve the current impasse. Finally, the article will further explore the key elements and ancillary safeguard measures for constructing such a statutory license system, aiming to provide beneficial theoretical references and institutional insights for improving copyright law regulation in the field of artificial intelligence in China and promoting the healthy and orderly development of AI technology within the rule of law.

## 2. Current Legal Controversies Regarding the Use of Works in AI Data Training

The development of generative artificial intelligence relies heavily on data training, through which patterns embedded in the training data are identified to construct the logical framework and knowledge architecture of large-scale models. The key issues arising at the research and development stage of generative AI are whether the use of copyrighted works in the course of data training falls within the scope of the copyright holder’s exclusive rights, and whether prior authorization from the rights holder is legally required for such use. Those issues, have become focal points of discussion in the current copyright field, yet academia and judicial practice are far from reaching a consensus on how to legally characterize and regulate it. This section aims to review the main controversial academic viewpoints surrounding this issue, laying a foundation for subsequent in-depth analysis.

### 2.1. The Theory of "non-expressive use"

Some scholars advocate that the use of works in the process of AI data training possesses significant characteristics of "non-specificity" or "non-expressiveness," and therefore should not be considered as the use of a work in the traditional sense of copyright law, nor should it fall within the scope of copyright control. This viewpoint is specifically elaborated as follows:

Firstly, from the perspective of the purpose and manner of use, when generative AI is undergoing data training, its intent is not to directly utilize the original expression of specific works to attract the public or satisfy their appreciative needs. Instead, it inputs massive quantities of works as raw data, and through algorithms, breaks down, extracts, and analyzes the inherent patterns, rules, or knowledge therein, such as linguistic structures and stylistic features, to serve the optimization of the model's own parameters and capabilities. In this process, the individualized expression of a single work is diluted, and the work more closely resembles a "consumed" raw material rather than an object to be appreciated or utilized independently.

Secondly, some proponents further argue from the perspective of balancing rights and incentive mechanisms that even if AI-generated outputs are identical or substantially similar in expression to pre-existing protected works, given that the holders’ right which can control over those of outputs is sufficient to guarantee their incentive to create, there is no need to grant them control over the data training processes involved in the production stage (“intput end”) [1].

As such inflation would not only potentially constitute an undue restriction on technological development and increase unnecessary transaction costs but it also goes against the fundamental goal of copyright law, which is to incentivize the creation of works and increase knowledge sharing. Therefore, defining the use of works in data training as "non-expressive use" helps to achieve a more

reasonable balance between encouraging technological innovation and protecting the rights and interests of authors in line with the long-term objective of encouraging scientific and technological progress and knowledge diffusion.

### 2.2. The Theory of "Fair Use"

Distinct from the "non-expressive use" theory, which attempts to exclude data training from the scope of copyright regulation, another group of scholars acknowledges that data training may constitute acts of use such as reproduction of works, but they argue that it can fall within the protective scope of the "fair use" doctrine, thereby being exempt from infringement liability.

The primary justification for this viewpoint is that the use of works in data training exhibits significant "transformative use" characteristics. Specifically, an AI model's reproduction of works is not for the simple purpose of reproducing the content of the works, but rather to learn from, analyze, and extract non-expressive elements (such as style, techniques, knowledge patterns, etc.) from them, thereby generating an AI model and output content with entirely new functions or purposes, capable of fulfilling diverse user commands [2]. This manner of use does not simply substitute the market for the original work but creates new value, which is highly consistent with the legislative spirit of the "fair use" doctrine that encourages innovation and promotes social progress.

Furthermore, considering industrial development and public interest from a value perspective, supporters believe that AI, as an emerging technology, relies on learning from massive amounts of data for its development. If every data training activity were strictly required to obtain prior authorization in the traditional sense, the high transaction costs, accumulation of license fees, and the potential unwillingness of some copyright holders to license would severely impede the development of AI technology. The "fair use" principle provides regulatory flexibility and innovation capacity for such new technologies of significant social value. Therefore, against the current backdrop of encouraging scientific and technological innovation and promoting overall social well-being, moderately applying "fair use" to exempt the use of works in AI data training, under the premise of meeting specific conditions, is conducive to fostering a more open and inclusive innovation ecosystem.

### 2.3. The Theory of "Authorized Licensing"

However, a considerable number of scholars adhere to the traditional stance of copyright protection, emphasizing the infringement risks of using others' works for data training without permission, and advocate that, in principle, prior authorization from copyright holders should be obtained.

This viewpoint holds that, firstly, when AI technology companies engage in data training—whether by acquiring data through web scraping, copying and organizing existing databases, or digitizing works and subsequently processing them with algorithms—they may all involve the direct exercise of exclusive rights such as the right of reproduction and the right of modification. These acts themselves fall within the regulatory scope of copyright law.

Besides, regarding the application of the "three-step test," large-scale, commercially motivated data training activities find it difficult to meet the elements of fair use [3]. First, the purpose of data training is often to develop AI models and services with commercial value, rather than for typical public interest or non-profit purposes generally recognized by international conventions and national copyright laws, such as personal study, research, commentary, or news reporting. Second, the quantity of works used in training is often massive, far exceeding the scope of “small and appropriate” use. Third, and most importantly, if AI models are allowed to generate content that competes with original works based on the unauthorized large-scale use of others' works, it is highly probable that this will unreasonably prejudice the normal exploitation of the original works and their potential market value, directly impacting the core interests of copyright holders. Therefore, these scholars maintain that obtaining permission from copyright holders is the fundamental way to safeguard their economic returns and moral rights (such as the right of attribution and the right to maintain the integrity of the

work), and it is also a necessary precondition for ensuring the healthy and orderly synergistic development of both AI technology and creation.

## 3. Copyright Infringement in the Context of Unlicensed Utilization of Third-Party Works for Data Training

Having reviewed the current main controversial academic viewpoints regarding the use of works in AI data training, this paper posits that both the "non-work-related use" theory and the "fair use" theory have certain theoretical limitations and cannot adequately address the challenges posed by technological development to copyright law. This paper advocates that, based on the fundamental principles of copyright law and in conjunction with the specific operational modalities of generative artificial intelligence, the act of using others' works without permission in the process of data training already constitutes an infringement upon the exclusive rights of copyright holders and usually does not meet the exemption conditions for fair use. A specific demonstration of this will be provided below.

### 3.1. The Act of Using Works in Data Training Constitutes "Use of a Work" in the Copyright Law Sense

To determine whether data training activities infringe copyright, it is first necessary to clarify whether they fall within the scope of "use of a work" as regulated by copyright law. This paper believes that, from a technical process perspective, both the data collection and subsequent machine learning stages inevitably involve the reproduction of works, which is one of the most central controlled acts in copyright law.

#### 3.1.1. Acts of Using Works in the Data Collection Stage

Using works for AI data training first involves the reproduction of works, which directly implicates the copyright holder's right of reproduction. The Chinese Copyright Law explicitly defines the right of reproduction as "the right to make one or more copies of the work by means of printing, photocopying, rubbings, sound recording, video recording, duplicating, rephotographing, digitizing, etc.". Prevailing copyright theories and practices in various countries indicate that for an act to constitute reproduction under copyright law, the work must be relatively stably fixed in a tangible form, thereby creating tangible copies of the work [4].

Data collection, as the initial stage and foundation of machine learning, has the core task of providing a sufficient quantity and diverse range of raw materials for model training. These materials themselves are often works protected by copyright law, such as various texts, images, and audio-visual materials [5]. Currently, AI technology companies primarily acquire such copyrighted works data through several typical pathways: first, by using web scraping technology to automatically capture and store publicly available works data from the internet. Specifically, developers use automated programs or scripts to simulate user or browser behavior according to predefined rules, sending web requests to website servers, receiving and parsing the returned web content, thereby achieving automatic crawling, extraction, and storage of information [6]; second, by directly copying large amounts of works aggregated in existing commercial or open-source databases; third, by digitizing non-electronic works, such as paper publications obtained through legal means, through processes like scanning and optical character recognition. The above-mentioned acts, regardless of their technical means, essentially involve the fixation, storage, and extraction of the works' information, forming digitized copies of the works, which are then stored on physical media such as hard drives to provide direct data input for subsequent machine learning [7]. Therefore, in the data collection stage, the use of works undoubtedly constitute reproduction under copyright law.

### 3.1.2. Acts of Using Works in the Machine Learning Stage

It is not only data collection but also the subsequent machine learning process that equally involves the further use of works. The effectiveness of machine learning largely determines the quality of the generated output. Meanwhile, AI progressively enhances its generative capabilities and intelligence level through in-depth analysis, pattern recognition, and algorithmic optimization of the collected data.

Specifically, data preprocessing, as a preparatory step for machine learning models, typically includes data cleansing (e.g., deduplication, error correction) and data annotation. In these processes, the original data often need to be reproduced again, edited, organized, or annotated to form formatted data or training sets suitable for machine learning. These processed and annotated data, as key elements in the training process, also constitute acts of reproduction under copyright law [8].

Model training is the core component of machine learning. During this stage, the model analyzes the preprocessed training data and continuously optimizes its parameters to ensure it can respond to user instructions more accurately. The essence of machine learning lies in the structured transformation and feature representation of information. By standardizing the acquired data, unstructured formats such as Word and Excel files are categorized and extracted to distinguish their information types, and ultimately converted into a machine-readable XML format. This transformation from human language to computer language is carried out according to rules set by programmers, with a one-to-one correspondence between the two. Therefore, this type of transformation still constitutes reproduction of the original work, and remains within the scope of acts under copyright law.

## 3.2. Unauthorized Data Training Does Not Constitute Fair Use

Having demonstrated that the act of data training constitutes reproduction of works, the next key question is whether this unauthorized act of reproduction can be defended by invoking the "fair use" doctrine. This paper argues that, based on the operational characteristics of machine learning and generative artificial intelligence, and in strict accordance with the standards of the "three-step test" in Chinese Copyright Law and relevant international treaties, the use of unauthorized works in data training usually cannot constitute fair use.

The latest revision of Chinese Copyright Law incorporates the "three-step test" as provided in international conventions and treaties such as the Berne Convention, the WIPO Copyright Treaty, and the TRIPS Agreement. This means that fair use must simultaneously satisfy: (1) it falls under specific circumstances explicitly listed by law or conforms to the legislative spirit; (2) the use of the work is confined to the necessary scope and does not affect the substantial part or core expression of the work; (3) it does not unreasonably prejudice the normal exploitation of the work, nor does it unreasonably prejudice the legitimate rights and interests of the copyright holder. However, current large-scale, commercial AI data training activities often find it difficult to satisfy these three conditions simultaneously [9].

Firstly, from the perspective of the purpose of use, the vast majority of data training is not for typical fair use purposes encouraged by the Copyright Law, such as personal study, research, classroom teaching, news reporting, or criticism of existing works. Its fundamental goal is often to train AI models with powerful content generation capabilities, which deviates from the public interest orientation of the fair use doctrine.

Secondly, from the perspective of the nature and quantity of use, AI models often need to be "fed" massive amounts of data to achieve ideal performance, far exceeding the requirement of "small and appropriate" quotation of works in traditional fair use. For example, in the case of AI-powered painting software, developers are required not only to input a substantial volume of visual artworks in order to train the AI in techniques such as compositional methods and color coordination, but also to utilize literary works to enable the AI to comprehend and process unstructured language [3]. Lastly, from the perspective of the impact on the potential market and value of the work. Large-scale

unauthorized data training activities can very easily constitute a substantial substitution for or potential threat to the normal exploitation of the original works. Thereby unreasonably harming the legitimate interests of copyright holders. To be more specific, if an AI model after learning from a large number of paintings in a specific style can easily generate new "works" that are highly similar or even indistinguishable from that style, it will undoubtedly impact the market sales and commercial value of the original painter's works [10]. As reports have already pointed out, a children's book titled Alice and Sparkle, created by a user within a few hours using AI tools such as ChatGPT and Midjourney, was published on Amazon. The book has been accused of extensive plagiarism, featuring low-quality images and illustrations with obvious inaccuracies [11]. In summary, unauthorized data training activities, due to the prevalence of their commercial purposes, the enormous quantity of works used, and their potential substitutive and detrimental effects on the market for original works, usually cannot pass the scrutiny of the "three-step test" and should not be deemed fair use. In summary, machine learning systems derive competitive advantages from large volumes of existing works, which may ultimately undermine the profits and prospects of the original authors. Thus, it is neither consistent with the definition of "the domain to be dominated by the public" nor a reasonable limitation on the "abuse of rights" by copyright holders.

## 4. The Legal Regulation Path for the Use of Works in AI Data Training: Intruduction of a Statutory License System

Against the backdrop where the preceding text has clearly demonstrated that unauthorized AI data training activities constitute copyright infringement and can hardly be exempted by applying the fair use doctrine, how to seek an effective legal regulation path that can both safeguard the basic rights and interests of copyright holders and promote the healthy development of the AI industry becomes a core issue urgently awaiting resolution. On the basis of fully considering the particularities of AI data training activities and the need to balance various interests, innovatively introducing a statutory license system is a preferred option that possesses both legitimacy and feasibility.

### 4.1. The Jurisprudential Basis for Applying a Statutory License System to Data Training

The statutory license system, as an important mechanism for limiting rights and coordinating interests within copyright law, has at its core the state's establishment of specific conditions through legislation, allowing users to use works in a legally prescribed manner when it is impossible or difficult to obtain prior permission from copyright holders, provided they pay statutory or reasonable remuneration. Introducing this system into the field of AI data training has a solid jurisprudential basis and practical considerations.

Firstly, the statutory license system aligns with the fundamental aims of copyright law to promote the dissemination of knowledge and technological progress. AI technology, as a significant driving force leading a new wave of technological revolution and industrial transformation, relies heavily on learning from and utilizing massive amounts of data. If the traditional model of individual authorization is rigidly adhered to, the high transaction costs will probably impede the pace of technological innovation. By setting uniform and convenient conditions for use, a statutory license can effectively reduce systemic transaction costs and ensure that AI developers can legally and efficiently utilize works data.

Secondly, the statutory license system can better balance the tension between the development of the AI industry and the protection of copyright holders' rights and interests. On the one hand, it acknowledges the use value of works in data training activities and, through a compulsory remuneration payment mechanism, ensures that copyright holders can receive reasonable economic returns from the reuse of their works. This, compared to completely excluding such use from copyright protection or simply categorizing it as fair use where rights holders receive no compensation, better reflects respect for and incentivization of creators' labor. On the other hand, it avoids the potential "stifling effect" on emerging industries that could result from an overemphasis

on the absolute control of exclusive rights. Allowing AI developers to use copyrighted works at an extremely low cost would deprive authors of adequate incentives and returns on their creative investment, thereby undermining the fundamental purpose of the copyright system [12]. Instead of that, the "permission to use but payment required" model embodies the dual value of "protection and limitation" in copyright law, aiming to achieve a dynamic balance between individual interests and the public interest.

In short, this paper proposes the addition of a statutory license for the scenario of "AI data training" to the existing statutory licenses specified under the current Copyright Law and the Regulations on the Protection of the Right of Communication through Information Networks. This will allow AI technology companies to use the works of others to train their models, subject to the condition of paying reasonable remuneration to the authors or rightsholders in accordance with the law, without having to seek individual authorization from each copyrights proprietor for every use. Moreover, the statutory licensing system enhances transactional efficiency by allowing use prior to compensation, thereby bypassing direct negotiations between copyright holders and users, and further promoting the dissemination and utilization of works [13].

### 4.2. Safeguard Mechanisms for the Effective Implementation of the Statutory License System

Having established the jurisprudential basis for applying a statutory license in the field of data training, what is more critical is to design a scientific, efficient, and operable system of implementation and safeguards to ensure that the system can truly be put into practice and achieve its intended purpose. This requires the collaborative efforts of legislators, judicial bodies, the industry, copyright collective management organizations, and other relevant parties. This paper will, in light of the implementation effects of existing policies, and put forward further recommendations to ensure the effective operation of the statutory licensing system.

#### 4.2.1. Establishing a Specialized Copyright Collective Management Organization for Data Training

Copyright collective management organizations (CMOs), as bridges connecting copyright holders and users of works, play an indispensable and critical role in the practical operation of a statutory license system. Such organizations are authorized by copyright holders to centrally manage licensing matters related to their works and to ensure that creators receive appropriate economic returns from their creations. In China, the five existing CMOs—including the Music Copyright Society and the Copyright Society of Written Works—have played a significant role in safeguarding the rights of authors and promoting the healthy development of the cultural industry.

With the advancement of AI technologies, data training has emerged as a critical area of application, relying on vast quantities of digitized works to optimize algorithms and enhance performance. In response to these developments, the establishment of a dedicated organization for the management of AI training datasets has become increasingly necessary. This organization would focus on copyright issues arising from data training and develop industry-specific rules and mechanisms to provide legally compliant solutions within an appropriate regulatory framework.

Those of existing CMOs have accumulated rich management experience in their respective areas, which can be referred to for the establishment and operation management for the new organization. Due to the uniqueness of data training, the new organization should have much stronger technical capabilities, work in more different domains, and have much more flexible modes of operation to support the rapidly developing AI industry. This will involve gathering resources from all aspects and building such a professional, technological, standardized collective management platform, which will be the cornerstone for effective running of the statutory license system in the data training field and making both the copyright owner and the user gain from this.

The core functions of such a specialized Data Training CMO should include, but not be limited to: First, establishing a convenient and efficient system for work registration and rights information

management. This involves encouraging and assisting copyright holders to voluntarily register their works, clarify rights ownership and licensing intentions, and utilize modern technological means to ensure the accuracy, completeness, and immutability of registered information, thereby providing a data foundation for subsequent license fee calculation and distribution. Second, being responsible for formulating and dynamically adjusting the royalty rates and distribution plans for statutory licenses in data training. This should comprehensively consider multiple factors such as the type of work, scale of use, commercial value of the model, stage of industry development, and the impact of market substitution on copyright holders, determining through transparent and fair procedures a payment standard that can both incentivize creation and promote technological development, and ensuring that license fees can be timely and accurately distributed to the relevant rights holders. Third, representing copyright holders in collective negotiations and contracting with AI enterprises, and supervising the implementation of the statutory license. Technical monitoring tools can be developed to track and record AI enterprises' use of works data, ensuring their compliance with all provisions of the statutory license and preventing abuse of rights.

Therefore, establishing a specialized Data Training CMO would not only provide effective solutions to the current challenges in copyright governance, but also lay a solid foundation for future technological advancement, thereby promoting overall cultural prosperity and technological progress. By integrating resources and enhancing cooperation and dialogue among stakeholders, it is possible to build a harmonious environment that serves the interests of both creators and users.

### 4.2.2. Improving the Regulatory System for AI Enterprises' Data Training Activities

In addition to leveraging the pivotal role of CMOs, constructing a robust external safeguard system that combines strong government regulation with industry self-discipline is equally crucial for standardizing the data training activities of AI enterprises and ensuring that the statutory license system is not abused.

Firstly, there should be enhanced transparency and disclosure obligations for AI enterprises. At present, China’s regulatory framework for AI-generated content is centered around the Interim Measures for the Management of Generative AI Services, which applies to all providers of generative AI technologies. The Measures are supplemented by the application of other relevant laws, regulations, and technical guidelines to further implement its specific provisions. Developers are required to enhance transparency and information disclosure in data training activities, ensuring that the use of works during the development and training of AI systems complies with applicable legal requirements. Regulatory authorities should explicitly require providers of AI services, especially developers of large models, to make necessary records, archives, and disclosures (without revealing core commercial secrets) regarding the sources, composition, scale of their training datasets, and the use of works. Violations should be penalized according to law to create an effective deterrent. This is not only fundamental to safeguarding copyright holders' right to know and right to supervise but also serves as an important basis for subsequent license fee calculations and dispute resolution. Consideration could be given to drawing on the principled provisions in the "Interim Measures for the Management of Generative AI Services" concerning the legality of data sources and information disclosure, and further refining and implementing them at the specific level of copyright protection.

Secondly, effective technical monitoring and compliance audit mechanisms should be established. Given the significant differences between AI-generated content and traditional forms of creation, the implementation and regulation of relevant measures face unique challenges and obstacles. For instance, traces of the use of copyrighted works in AI training are often easily concealed or obscured, and malicious developers may exploit this technical feature to erase records of unauthorized use. However, blockchain technology offers a potential solution by preserving such traces in a tamper-proof manner. Through cryptographic techniques, blockchain can secure training data in block structures, ensuring that the records are immutable and verifiable, thereby enabling retrospective audits of data usage [12].

Furthermore, industry self-discipline and standard-setting should be promoted. AI industry associations and other organizations should be encouraged to play an active role in organizing the formulation of industry standards, best practice guidelines, and ethical codes related to data training, guiding enterprises to consciously comply with laws and regulations and respect intellectual property in data acquisition, use, and annotation. A healthy and orderly industrial development ecosystem should be fostered by establishing inter-enterprise trust mechanisms, user complaint handling and feedback channels, and positive incentives for compliant enterprises.

It will guarantee that the right type of statutory license system for the AI data training is effective in its implementation and genuinely safeguards the legitimate rights of the copyright holders while also not diminishing the motivator force for innovation. Rather, it should lead to the harmonious progress of technology, culture, and law.

## 5. Conclusion

In conclusion, the development of AI must be guided and regulated by sound legal and policy frameworks. With respect to the use of copyrighted works for data training, this article adheres to a human-centered governance principle, arguing that only by striking a balance between the interests of copyright holders and AI technology providers can a constructive interaction among technology, art, and law be achieved, fostering their coordinated development.

Such a balance involves not only the distribution of economic benefits, but also the protection of creative incentives and the preservation of cultural diversity. In this process, the law plays a fundamental role in providing a clear framework of action for all parties involved, ensuring that the advancement of AI does not come at the expense of individual rights, while also guaranteeing that technological progress benefits a broader segment of society.

Policymakers must thoughtfully consider how to safeguard copyright protection while also creating the necessary legal space to support AI development. This is essential to achieving a harmonious integration of technological innovation and humanistic values, and to promoting broader social progress.

This article advocates for the implementation of a statutory licensing system in the field of data training as a viable solution to reconcile competing interests. Such a system can equitably balance both sides: on one hand, it permits AI companies to make broad use of existing works; on the other hand, it ensures that copyright holders' rights are fully protected. This approach aligns with the dual objectives of copyright law—to protect and to reasonably limit rights—thus reflecting its underlying value orientation.

## References


[1] X. C. Liu, "Non-Work-Related Use" and its Legitimacy Justification in Generative Artificial Intelligence Data Training, Law Forum (3) (2024).

[2] J. Y. Zhang, S. F. Wang, Research on Copyright Fair Use in Large Model Data Training, Journal of East China University of Political Science and Law (4) (2024).

[3] Q. W. Li, Legal Regulation Path for the Use of Works in Algorithm Training under Copyright Law, Science and Technology and Publishing (7) (2024).

[4] Q. Wang, C. Chu, A preliminary exploration of the boundary between artificial intelligence and copyright: Legal challenges and thoughts under technological progress, Chinese Editor (8) (2024) 58.

[5] M. Kretschmer, T. Margoni, P. Oruç, Copyright law and the lifecycle of machine learning models, IIC-International Review of Intellectual Property and Competition Law 55(1) (2024) 110-138.

[6] B. Massimino, Accessing online data: Web-crawling and information-scraping techniques to automate the assembly of research data, Journal of Business Logistics 37(1) (2016) 34-42.

[7] C. M. Dahl, T. S. Johansen, E. N. Sørensen, C. E. Westermann, S. F. Wittrock, Applications of machine learning in document digitisation, arXiv preprint arXiv:2102.03239 (2021).

[8] C. Zong, R. Xia, J. Zhang, Data annotation and preprocessing, in: Text Data Mining, Springer Singapore, Singapore, (2021) pp. 15-31.

[9] T. Zhang, Legal risks and inclusive and prudent regulation of generative artificial intelligence training datasets, Comparative Law Review (4) (2024) 92.

[10] Y. Gao, Regulation of copyright infringement by artificial intelligence training data, China Publishing Journal (5) (2024) 14.

[11] A. Levendowski, How copyright law can fix artificial intelligence's implicit bias problem, Wash. L. Rev. 93 (2018) 579-630.

[12] Y. H. Liu, Y. S. Wei, The copyright infringement problem of machine learning and its solution, Journal of East China University of Political Science and Law (2) (2019) 76.

[13] Y. Gao, D. Y. Hu, Challenges and Responses of Machine Learning to the Copyright Fair Use System, Electronic Intellectual Property (10) (2020).