Research on Copyright Fair Use in Generative AI Training Phase

Remila Taximaimaiti
School of Law, China Jiliang University, Hangzhou, Zhejiang, China

Abstract: In the era of rapid advancement in generative AI technology, the growing demand for massive datasets during training phases poses new challenges to copyright systems. This paper explores the establishment of fair use mechanisms for generative AI training stages. It begins by analyzing the technical principles of generative AI and their implications for copyright law, then examines the limitations of the current "three-step test" framework in adapting to generative AI development. Finally, the study highlights the critical role of fair use provisions in fostering technological innovation, cultural prosperity, and social progress, advocating for specialized exceptions regarding fair use in generative AI data training to reconcile diverging interests among stakeholders.

Keywords: Generative AI; Copyright; Copyright infringement; Data; Fair Use

1. Introduction

The rapid advancement of intelligent technologies has made machine learning a focal point of public attention. At its core, machine learning involves training models through massive datasets to enhance algorithm performance. During the training phase of generative AI, extensive data collection and integration into algorithmic frameworks create vast information repositories. Through iterative learning and optimization processes, these systems ultimately generate content with distinctive innovative features. However, copyright concerns surrounding training datasets have raised legal challenges. Current copyright law provisions on fair use limitations struggle to adequately address learning behaviors during AI training, while criminal liability for copyright infringement would increase development costs and hinder industry progress, sparking legal disputes [1]. The official launch of China's ChatGPT Wenxin Yiyang on October 17, 2023 marked significant technological breakthroughs in generative AI. Concurrently, copyright-related controversies have intensified and attracted growing judicial attention. Properly addressing copyright protection during AI training remains crucial. This study examines the legal validity of fair use mechanisms in generative AI training processes and explores framework development strategies. To resolve these issues, comprehensive analysis is required to establish a balanced solution that resolves copyright conflicts while fostering AI industry growth.

2. The Preparatory Issues of Fair Use of Copyright

2.1 Definition of Generative Artificial Intelligence

Generative artificial intelligence is a technology that generates new content by leveraging existing data patterns. The workflow of generative AI can be divided into three stages. First is the machine learning phase, where the system analyzes and processes input data during design optimization. Next comes the algorithm execution phase, where operators input parameters and adjust model specifications to perform computations. Finally, the result presentation phase completes the generation of final content after computational processing [1]. Through extensive data training, generative models learn to recognize distribution patterns in input information, thereby understanding data characteristics and statistical patterns. These data formats include images, text, audio, and more. By analyzing core data patterns, the models create new content. After understanding data distributions, they generate new data through random methods or predefined rules. The generated content retains statistical features and structures similar to the training dataset, yet typically represents entirely original creations that have never existed before.

With technological advancements, generative artificial intelligence has sparked a series of legal challenges. Scholars argue that training databases not only constitute a core component of generative AI systems but also serve as a collective legal interest entity, providing a critical regulatory focus for governing such technologies. However, the

development of training datasets carries legal risks associated with unauthorized use of copyrighted materials. Striking a balance between technological innovation and copyright holders' rights has become an unavoidable key issue in this context.

2.2 The Challenge of Generative Artificial Intelligence to Copyright System

If AI-generated content is recognized as intellectual property, copyright ownership issues inevitably arise [2]. When AI-generated materials infringe upon others 'copyrights, challenges emerge regarding liability determination and infringement accountability. These questions constitute critical challenges during the output phase of generative AI systems, with extensive academic analysis addressing these issues. The copyright concerns discussed in this paper during the front-end training phase represent the most pressing and practical challenges in AI copyright protection. At this stage, machine learning not only initiates infringement activities but also becomes the most contentious aspect of infringement determination. Effectively resolving front-end challenges can significantly reduce algorithmic infringement risks during mid-phase operations [3]. The massive data datasets required for AI training form the foundation of technological advancement, where insufficient data supply substantially impacts training efficacy and performance. Concurrently, AI development inherently involves extensive use of human-created works. Regarding this complex copyright governance issue, mainstream perspectives exhibit two starkly contrasting stances [4]. On one hand, it is essential to ensure AI technology development does not compromise creators' rights, on the other hand, adequate space must be provided to avoid stifling technological innovation and industrial progress. Finding a balance between copyright protection and AI technological advancement thus emerges as a shared challenge for policymakers and society at large.

3. Current Status of Research on Reasonable Use Rules

3.1 Regulations on Fair Use Rules at Home and Abroad

The Copyright Law protects authors 'rights to encourage creative activities and foster the creation of outstanding works. By defining the rights of distributors, it promotes the widespread dissemination of works, ultimately serving broader social interests. The Berne Convention further stipulates that member states may enact laws permitting reproduction of works without author consent under exceptionally rare circumstances, provided three conditions are met: such reproduction must not be widespread or mass-produced but limited to specific cases, it must not compete with the author's legitimate income-generating activities (e.g., mass reproduction of textbooks to replace legitimate purchases), and it must genuinely respect the author's economic interests and moral rights, avoiding unjust losses under the guise of "exceptions" [5].

While all countries establish copyright restrictions and exceptions, there are significant differences in terminology and legal frameworks. The U.S. Copyright Act introduces the "fair use" principle but does not specify detailed scenarios qualifying as fair use. In civil law jurisdictions, copyright laws classify unauthorized use of works under "exceptions and limitations of rights," typically listing various specific exceptions while often lacking detailed explanations of the underlying principles. Consequently, when encountering non-exempt cases, judges may face challenges in determining whether such acts constitute infringement [6].

In the "Google Digital Library Case," the U.S. Court of Appeals for the Second Circuit ruled that Google's actions constituted fair use. The company's reproduction of books served search functionality, constituting transformative use. The displayed excerpts were both minimal in quantity and essential for the purpose. Moreover, previewing fragments did not replace the complete works but actually contributed to boosting original sales [7].

The copyright system is designed to protect authors 'rights, while the fair use doctrine seeks to balance copyright holders' interests with users' needs by permitting limited use of works without authorization or compensation under specific circumstances. This mechanism facilitates knowledge dissemination and cultural development. The flexible nature of fair use provisions enables adaptation to technological advancements and social changes, ensuring copyright laws remain relevant to evolving communication technologies and usage patterns.

China has established a "three-step test" framework for fair use, comprising three essential criteria: first, the use must be confined to specific contexts, second, it must not impair the normal use of the

work, and third, it must not unduly infringe upon the legitimate rights of the rights holder. In judicial practice, determining whether an act constitutes fair use typically involves comprehensive evaluation of multiple factors, including the intent and methods of utilizing another work, the genre and distinctive features of the original work, the proportionality and substantive significance of the referenced content, as well as potential impacts on the market potential or economic value of the original work [8].

In the Wang Xin v. Google case, Google displayed scanned books to users in "fragmented" formats while providing search results related to keywords. This practice fell under the category of "right of communication through information networks," and the court ruled it constituted fair use. However, regarding Google's practice of "scanning entire books," the court did not consider it to constitute fair use under the "right of reproduction."

3.2 Research on Reasonable Use of Copyright Applicable to Generative AI Training

Some scholars argue that training large-scale models does not generate independent replicas nor attempt to replace original works in the market. The primary objective is to enhance content creation efficiency, which justifies exemption from copyright infringement liability. The various property rights stipulated by copyright law essentially reflect institutional patterns in market transactions, indicating that legislation establishes stable revenue mechanisms for copyright holders. Unauthorized use of copyrighted works that substitutes existing market demand may adversely affect rights holders' income, thereby diminishing their creative motivation. Therefore, maintaining the original market structure and interests of rights holders is fundamentally justified [9].

Zhang Jinping argues that legal protection for works is not inherently granted. On the contrary, modern copyright law provides legal safeguards precisely to encourage authors to create more works, thereby achieving public policies that promote social, economic, and cultural prosperity [10]. Scholars advocating non-copyright usage distinguish between "work usage" and "data usage," asserting that artificial intelligence technology essentially utilizes underutilized resources—data—left after works are disseminated online, rather than directly accessing or appreciating copyrighted content. Thus, they contend that the core of copyright remains unviolated.

Wu Handong proposed that text and data mining should be incorporated into the fair use framework based on the adoption of the "three-step test method" [11]. Wan Yong suggested adding a dedicated fair use exception clause in the revised Copyright Law Implementation Regulations, with its scope strictly limited to data mining and restricted to acts of reproduction, storage, and public dissemination. Regarding the definition of responsible parties, he recommended adopting China's practical circumstances without imposing restrictions [12].

In the new landscape of production relations, how to achieve rational allocation of interests through the definition of rights and their boundaries remains a core issue requiring in-depth attention. The confrontation between AI companies and copyright holders will hinder the healthy development of AI technology and industries. The efficient growth of AI relies on continuously updated and expanded textual resources as foundational support, which play a critical role during the learning phase of artificial intelligence. High-quality and diverse training data are essential for ensuring the quality of outputs from large model-based generative AI systems, reducing discrimination and biases, and maintaining content and cultural diversity. Without active participation and contributions from creative communities, the AI field will struggle to sustain long-term innovation momentum and disruptive advancements.

Scholars who hold a negative stance argue that while applying fair use rules to AI-generated content and adopting the Japanese model may seem appropriate at present, from a long-term perspective, as copyright licensing mechanisms continue to mature, it is essential to revert to or shift toward statutory licensing approaches. This would better facilitate the balance of interests among relevant stakeholders.

Jiang Yike argues that unlike the fair use doctrine which aims to promote public interest by excluding market transactions, the statutory licensing system essentially serves as a supplement or alternative to free market mechanisms. By protecting copyright holders 'legitimate rights while permitting work usage under specific conditions, it achieves the public interest goal of facilitating knowledge and information dissemination to the public. This mechanism not only respects copyright holders' rights but also meets society's demand for knowledge sharing [13]. Liu Youhua and Wei Yuanshan point out that the fair use doctrine tends

to drive technological advancement in machine learning fields, whereas the statutory licensing system focuses more on balancing stakeholder interests—particularly demonstrating stronger protection for copyright holders—thereby achieving equilibrium between developers and rights holders. However, the implementation of statutory licensing systems relies on technological and institutional support, which remains imperfect at present. Consequently, it is regarded as an idealized solution requiring time-tested validation and practical adjustments.

Xiong Qi argues that copyright holders and users typically determine licensing terms through free negotiation, with such agreements reflecting both parties 'bargaining power based on equal status. However, when prior negotiations create transactional barriers, statutory licensing systems intervene to restrict copyright holders' decision-making authority and pricing power. Under statutory licensing, users may legally obtain usage rights after paying specified fees. This mechanism aims to ensure public access to intellectual creations for groups otherwise unable to obtain them under normal circumstances, thereby serving public interests [14].

4. Research Prospect on Copyright Fair Use in Generative AI Training Phase

Regarding research on fair use of copyright during generative AI training phases, there currently exist two starkly contrasting viewpoints. On one hand, proponents argue that to drive technological progress and innovation, legal leniency should be granted for AI systems utilizing human-created works. They advocate permitting reasonable use of copyrighted materials within defined parameters to facilitate AI development and application. This perspective emphasizes technological innovation's significance, asserting that appropriate usage can benefit society as a whole by advancing education, scientific research, and cultural development. On the other hand, critics emphasize strict copyright protection, contending that unauthorized use of copyrighted works may infringe upon original authors' legitimate rights, undermine creators' motivation mechanisms, and negatively impact cultural innovation and intellectual property safeguards. This stance maintains that even AI technological advancement must be conducted while respecting and protecting copyright principles.

With the rapid advancement of generative AI technology, the tension between its massive data requirements during training and copyright fair use rules has become increasingly prominent. Training generative AI necessitates access to vast datasets that inevitably involve utilization of existing works. Current copyright law provisions on fair use prove inadequate in addressing this emerging challenge, particularly within the framework of the "three-step test." Critical issues such as interpreting "specific circumstances," ensuring "no disruption to normal use of works," and preventing "unreasonable infringement of rights holders' legitimate interests" have emerged as pressing challenges requiring immediate resolution.

Given the massive data volume required for training generative AI models, implementing licensing mechanisms may create operational pressures during training processes, potentially disrupting the balance of interests between users and copyright holders. Additionally, the inherent opacity of generative AI in data collection processes could lead to market regulation mechanisms failing. Drawing from historical interactions between technological advancements and fair use frameworks, establishing specialized fair use rules tailored for generative AI model training appears to be a viable approach. This strategy would address current licensing model conflicts, facilitate continuous AI evolution, and rebalance stakeholder interests. The author emphasizes that fair use policies hold significant implications for technological innovation, cultural prosperity, and social progress. When revising copyright law implementation rules, introducing specific exceptions for generative AI data training could effectively resolve disputes among stakeholders. These strategies aim to strike a balance between copyright protection and AI advancement, creating a flexible legal framework that fosters harmonious development between technology and legal systems.

References

[1] Wu Handong: General Theory of Intellectual Property Rights, Beijing: Renmin University of China Press, 2020

[2] Wang Qian: Copyright Law, Beijing: Renmin University of China Press, 2015.

[3] Cai Yuanzhen. The applicability basis and rule construction of statutory licensing for machine learning copyright. Intellectual Property, 2024, (11):77-93.

[4] Sun Jingzhou. Copyright Dilemma and Solutions in AI Training: An Analysis of Modular Licensing Mechanisms. Intellectual

Property, 2024, (11):94-111.
[5] Li An: "Copyright Law Analysis of Machine Learning Works: Non-Work Use, Fair Use and Infringement Use", in Electronic Intellectual Property, No.6, 2020, pp. 63-67.
[6] Zhang Tao. Legal Risks and Inclusive Prudent Regulation of Generative AI Training Data Sets. Comparative Law Studies, 2024, (04):86-103.
[7] See Agreement on Guidelines for Classroom Copying in Not-For-Profit Educational Institutions with Respect to Books and Periodicals, Published in House Report 94-1476, 1976.
[8] Lin Xiuqin: "Reconstructing the Fair Use System for Copyright in the Era of Artificial Intelligence," in Law Research, No.6, 2021, p. 183.
[9] Xu Xiaoben, Yang Yinan: "On the Fair Use of Copyright in Artificial Intelligence Deep Learning," Journal of Jiaotong University Law, 2019, No.3, p. 35.
[10] Zhang Jinping. The Dilemma of Fair Use in AI Works and Its Solutions. Global Legal Review, 2019,41(03):120-132.
[11] Wan Yong: "The Dilemma and Solutions of the Fair Use System in Copyright Law in the Era of Artificial Intelligence", Social Sciences Journal, No.5, 2021.
[12] Wu Handong: "Copyright Law Issues Regarding AI-Generated Works", China Foreign Law Review, No.3, 2020.
[13] Jiang Yike: "Exploration of Digital Music Copyright Licensing Models: On the Necessity of Statutory Licensing and Its Institutional Framework," East China Law Review, 2019, No.1, p. 154.
[14] Xiong Qi: "Decoding the Legitimacy of Statutory Copyright Licensing and Institutional Substitution", in Intellectual Property, No.6, 2011, p. 39.

