Proceeding of ICGPSH 2025 Symposium: Exploring Spelling Evolution and Educational Approaches to Dyslexia
DOI: 10.54254/2753-7048/2025.HT25556

Establishing a Statutory Licensing Framework for Data Mining in Generative AI

Yifan Xue

College of Law, Yanbian University, Yanji, China
X18101488129@outlook.com

Abstract. The rapid development of generative artificial intelligence (AI) and its expanding commercial applications have intensified the demand for large-scale, high-quality datasets, leading to frequent copyright disputes. These conflicts reveal the limitations of the current copyright framework in addressing the legality of text and data mining (TDM) during model training. Focusing on the use of commercially non-public, copyright-protected content in the training phase, this paper analyzes the tension between technical necessity and the legal constraints arising from the expansion of reproduction and adaptation rights. It examines the shortcomings of existing Chinese regulations—including the Administrative Measures for Generative Artificial Intelligence Services—in providing an operable and scalable authorization mechanism. Drawing on the historical functions of copyright and a comparative analysis of TDM exemption regimes in the EU and Japan, the study proposes a Chinese model based on a statutory licensing regime tailored to TDM. The proposed legal framework employs a semi-open legislative technique, featuring a "general clause–illustrative list–residual provision" structure, and is supported by the establishment of a new type of copyright management organization endowed with public authorization. This institution would oversee registration, remuneration distribution, and rights verification, ensuring lawful usage without prior consent while enabling equitable compensation. The proposed system aims to overcome licensing inefficiencies, reduce transaction costs, and promote coordinated development between copyright protection and data governance in the digital economy.

Keywords: Generative Artificial Intelligence, Text and Data Mining, Copyright Law, Statutory Licensing, Copyright Management Organization

1. Introduction

The rapid advancement of artificial intelligence technologies has led to the widespread application of generative AI in commercial domains, with model training increasingly reliant on massive datasets. However, the tension between the high cost of copyright licensing and the growing demand for large-scale data has triggered frequent legal disputes, highlighting the urgent need to address copyright issues in this context. For instance, in Butterick v. GitHub, the plaintiffs—a group of software developers—alleged that the defendants infringed copyright by using protected source code without authorization to train Codex and Copilot [1]. Similarly, Getty Images v. Stability AI

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2025.HT25556

involved accusations of unauthorized copying of millions of copyrighted images and associated metadata to train Stable Diffusion [2]. In The New York Times v. OpenAI and Microsoft, the plaintiffs claimed that the defendants committed direct copyright infringement by using a large volume of news articles to train ChatGPT without permission [3]. These cases reflect the unresolved legal challenges that generative AI systems face during the training phase. Striking a balance between protecting the interests of copyright holders and fostering technological innovation has become a pressing regulatory concern.

Various academic proposals have been advanced to address the legitimacy of using copyright-protected content in generative AI training. Some scholars advocate adjusting the doctrine of fair use, drawing on models such as the EU's Directive on Copyright in the Digital Single Market (DSM Directive) [4] or the more flexible fair use regime in the United States [5]. Others propose statutory or compulsory licensing schemes as a solution [6]. Nevertheless, no academic consensus has emerged, and these unresolved issues affect not only the practical legality of AI training but also the theoretical development of copyright law and its future adaptability.

This study focuses specifically on the copyright implications of using commercially non-public, protected content in the training of generative AI models. It examines the conflict between the technical necessity of large-scale data extraction and the expanding scope of copyright protections. Grounded in the origins of copyright as a mechanism for market allocation and commercial incentive, the study explores the legitimacy and viability of a statutory licensing regime as a legal response to current regulatory deficiencies.

2. The copyright dilemma in the use of protected works for generative AI training

2.1. Technical imperative: the role of large-scale data and the risk of data deficiency

Generative AI relies heavily on vast and diverse datasets. Whether in natural language processing, image generation, or multimodal tasks, the underlying algorithms require deep learning models to extract patterns from large-scale inputs. In a 2024 submission to the UK Parliament, OpenAI stated that without access to copyrighted content, its tools would "not function." [7] Models such as GPT-3 are trained on corpora containing hundreds of billions of words, and research has shown that models with larger parameter counts exhibit stronger meta-learning capabilities [8,9]. Open datasets like MINT-1T have been demonstrated to significantly enhance inference performance in systems such as XGen-MM [10]. According to the China AI Development Report 2024, large models are evolving toward a "data-centric" paradigm in which data plays a central role in driving performance [11].

However, if copyright law requires that all training data be authorized by rightsholders, this would significantly raise barriers to use and increase the risk of biased datasets. The Interim Measures for the Administration of Generative AI Services in China mandate that training data be properly licensed, which may lead to the exclusion of marginalized groups from training samples and reduce the social representativeness of AI systems. For instance, in the collection of health-related data, individuals with higher education or income levels are more likely to consent to data use, while elderly individuals and those with chronic illnesses tend to opt out [12]. This results in sample distortion, increasing the risk of AI misjudgment and reinforcing structural inequality. Therefore, legal frameworks should be designed with both legitimacy and inclusiveness in mind, ensuring that compliance mechanisms do not result in technological discrimination. This approach aligns with Article 1 of China's Copyright Law, which emphasizes the promotion of cultural and scientific development [13].

DOI: 10.54254/2753-7048/2025.HT25556

2.2. Legal risks in data mining: from rights expansion to institutional inflexibility

2.2.1. Conceptual boundaries between copyright and data mining

Data encompasses far more than copyrightable works. Only when data exhibits originality and a specific form of expression does it fall within the legal definition of a "work" under the Copyright Law. Accordingly, the "copyright issues of generative AI training data" discussed here do not concern all data indiscriminately, but specifically refer to the use—during training—of content protected by copyright, such as literary works, images, music, or videos. Not all data mining activities fall within the purview of copyright regulation; only those involving "works" are subject to such legal scrutiny. Thus, the core legal issue centers on whether and how copyright-protected works may be lawfully used in algorithmic training, rather than whether "data" in general may be used.

2.2.2. Expansion of reproduction and adaptation rights

Under the current Copyright Law, the scope of reproduction and adaptation rights has expanded in response to technological developments, placing data mining—especially for model training—under considerable legal risk. Digitization is now explicitly included within the scope of reproduction, meaning that web scraping and text conversion activities are subject to regulatory control [13].In practice, if a company collects protected books, images, or audiovisual content without authorization —even solely for training purposes—it may be held liable for direct infringement. During pre-processing, such materials must be downloaded, archived, transcoded, or reformatted, all of which constitute acts of reproduction under the law.

Moreover, structural transformation and semantic extraction from protected works—such as vectorization, keyword extraction, or stylistic normalization—may be construed as the creation of derivative works, thereby implicating adaptation rights. When large volumes of content are classified, annotated, and compiled into a unified training corpus, issues related to compilation rights may also arise. While certain interim reproductions may be exempt under the Regulations on the Protection of Computer Software and the Regulations on the Right of Communication Through Information Networks, these exemptions apply only when no substantive transformation of the original work occurs—an unlikely condition in the context of AI data mining, which typically entails reformatting and semantic reconstruction. Furthermore, China's current system of copyright exceptions does not adequately accommodate the broad scope of text and data mining (TDM) activities, leaving such applications in a legally precarious position [14].

2.2.3. Regulatory limitations of the administrative measures for generative AI

The 2023 Interim Measures for the Administration of Generative AI Services address certain copyright issues but remain institutionally constrained. Article 7 requires that training data consist of "copyrighted works legally obtained," thereby establishing a permission-based mechanism that is highly inefficient at scale. For example, acquiring licenses for tens of millions of texts to train a language model entails prohibitive transaction costs in terms of time and labor—far exceeding the capacity of most startups. This inefficiency contradicts the fundamental goal of copyright law: to encourage innovation and the dissemination of knowledge. As emphasized in the TRIPS Agreement, intellectual property protection should promote technological innovation, facilitate technology transfer, and strike a balance between the rights of holders and users in order to enhance social and

DOI: 10.54254/2753-7048/2025.HT25556

economic welfare [15]. Additionally, while the Measures prohibit the inclusion of infringing content in training datasets, such provisions are difficult to enforce. For instance, collecting publicly available content via web crawlers or using legally acquired material may still blur the line between lawful and infringing use [16].

Several alternative mechanisms have been proposed in academic discourse. Although some scholars hoped that the doctrine of fair use could offer a solution, as David Atkinson notes, allowing generative AI to copy vast volumes of copyrighted works under fair use would logically entail granting the same privilege to humans—thereby undermining the economic foundation of the copyright system [17]. Moreover, China's judicial practice displays inconsistent standards regarding fair use, characterized by divergent rulings and a lack of clear precedent [18]. Other scholars have suggested narrowing the scope of reproduction rights. However, Article 10(5) of China's Copyright Law explicitly includes digitization within the definition of reproduction, rendering such a move legally precarious. More critically, many existing proposals focus on weakening rights without addressing the structural issue of how knowledge interests are institutionally distributed. Copyright historically emerged as a mechanism for allocating knowledge-related value. Returning to these historical foundations may provide a normative basis for integrating text and data mining (TDM) into the modern copyright system.

3. Historical analysis: the case for a statutory licensing regime

3.1. The role of merchants in the origin of copyright law

To address the current regulatory dilemma surrounding the application of copyright exceptions to text and data mining (TDM), it is not sufficient to rely solely on comparative analyses of contemporary foreign legislation. Like other legal institutions, copyright was shaped by the interests and power structures of its time. Reexamining its historical origins helps to reveal which groups the system was initially designed to benefit and provides a more grounded basis for reassessing its legitimacy in the age of artificial intelligence.

Take the Statute of Anne (1710) as an example. Enacted after the collapse of the royal printing monopoly, this statute was designed to help publishers re-establish market control through statutory copyright. While the law nominally recognized authors as the initial rightsholders, in practice it functioned to secure publishers' ownership. As long as authors assigned their rights, the law enabled booksellers to reclaim market privileges that had been lost under the earlier licensing regime [19]. Scholars have argued that the Statute of Anne is more accurately viewed as a piece of trade regulation than as legislation protecting creative labor [20]. A similar pattern emerged in France: in 1777, King Louis XVI issued a decree recognizing authors' publishing rights, but the real aim was to resolve conflicts between Parisian and provincial publishers. The assertion of "authorial ownership" served primarily to reinforce the exclusive trading rights of Paris-based firms [21].

Independent historical research by Chinese scholar Huang Haifeng further supports this commercial logic: "Merchants have always been both beneficiaries and drivers of copyright law's development. The formation of U.S. copyright law, for example, was driven by lobbying from business groups with direct interests in copyright." [22] In short, copyright was originally designed as a legal mechanism to transform knowledge into tradable capital. While some creators have benefited from this structure, individual authors—especially those not operating professionally—have rarely constituted an effective lobbying force [23]. As noted in the British Commission on Intellectual Property Rights' report Integrating Intellectual Property and Development Policy, "Intellectual property is now widely seen as an economic and commercial right. Most holders are

DOI: 10.54254/2753-7048/2025.HT25556

corporate entities, not individual creators." [24] Therefore, copyright law has historically functioned to facilitate the marketization of knowledge—an institutional role that remains highly relevant today.

3.2. Technological change as a catalyst for copyright innovation

In the modern era, the scope of copyright protection has expanded from books to encompass broadcasting, phonograms, and digital platforms. Yet, the core logic of the system has remained consistent: it serves the structural interests of industrial capital. As Norio Watanabe has observed, the patent system functions as a form of industrial policy by creating exclusive rights that delay market entry by non-patent holders and thereby artificially constrain total output [25]. Copyright operates in a similar fashion—constructing legal scarcity around knowledge-based goods to secure market advantages for rights holders. The background of the U.S. Copyright Act of 1909 and its intersection with the rise of broadcasting exemplify this logic: both systems institutionalize exclusivity not simply to protect creators, but to shape markets and control access to intellectual resources.

The U.S. Copyright Act of 1909 provides a compelling case in point. This legislation introduced statutory rights over the mechanical reproduction of musical works. Yet, in response to the monopolistic practices of the Aeolian Company—which dominated the market for piano rolls—the law simultaneously introduced a compulsory licensing provision. Under this provision, once a copyright holder authorized one party to reproduce a musical work, all others were entitled to make similar reproductions upon payment of a statutory royalty (set at two cents per copy). This approach curbed monopolistic behavior while balancing the interests of rights holders and users [26].

In the early 20th century, the rapid rise of radio broadcasting transformed it into the dominant medium for mass communication. Major networks such as NBC and CBS regularly broadcast copyrighted music without obtaining authorization, triggering large-scale infringement disputes with composers and music publishers. In response, rights holders established the American Society of Composers, Authors and Publishers (ASCAP), which began filing lawsuits against broadcasters and demanding compensation. In 1931, the courts ruled that radio transmissions constituted "public performance," thereby affirming the applicability of copyright law to radio use.

However, this legal recognition did not resolve the growing transaction costs of licensing. Negotiating individual permissions with thousands of broadcasters proved prohibitively expensive and administratively unworkable. By the mid-1930s, ASCAP and the radio networks reached a compromise: ASCAP would offer blanket licenses covering its entire repertoire in exchange for a standardized royalty fee, to be distributed among rights holders. This arrangement did not emerge out of concern for individual authorship, but rather as a systemic response to industrial conflict and legal impasse. It institutionalized a new mechanism—collective licensing—not to protect creativity per se, but to restore predictability and order in a chaotic rights environment. The system absorbed technological disruption by reconfiguring rights enforcement in a way that prioritized market operability over authorial control, thereby reaffirming copyright's role as a facilitator of trade rather than a pure moral entitlement.

4. Comparative study of foreign approaches to copyright and generative AI

In the comparative analysis of foreign regulatory approaches to copyright and generative AI, this paper focuses on the European Union and Japan—two civil law jurisdictions with institutional structures and legislative philosophies comparable to those of China. This similarity enhances the relevance and applicability of their regulatory experiences in the Chinese context. In contrast,

DOI: 10.54254/2753-7048/2025.HT25556

common law countries such as the United States and the United Kingdom rely primarily on judicial precedent and market-driven governance, diverging significantly from China in legal methodology and institutional configuration. As such, they are not the primary focus of this chapter.

4.1. Copyright exemptions and data intermediaries in the EU framework

4.1.1. Legislative design of the Data Governance Act

The European Union's Data Governance Act (DGA), which came into effect in 2023, represents a cornerstone of the broader European Data Strategy. Its principal aim is to reconcile privacy protection with enhanced data sharing to foster digital innovation and economic growth [27]. Rather than mandating data sharing, the DGA adopts a "voluntary–compliant–intermediated" approach and establishes three foundational systems: a reuse mechanism for public-sector data, a registration regime for data intermediation services, and a framework for "data altruism." [28] In particular, intermediary platforms are tasked with coordinating authorization procedures and establishing transaction protocols, thereby reducing access barriers while safeguarding associated risks.

Conceptually, the DGA is grounded in the principle of voluntary authorization, emphasizing the autonomy of data controllers and data subjects. It thus reflects a "trust-based intermediation" model that differs from statutory licensing systems, which are typically driven by public interest mandates. Although the DGA does not introduce compulsory use or remuneration schemes in the copyright sense, its institutional design—especially the establishment of neutral data intermediaries—offers valuable lessons for China. Given the lack of a unified data authorization mechanism in China, the DGA provides a reference model that combines administrative oversight with market-based intermediation, potentially bridging the gap between lawful data access and rights protection.

4.1.2. TDM exemptions and their limitations under the DSM directive

Articles 3 and 4 of the European Union's Directive on Copyright in the Digital Single Market (DSM Directive) establish specific exceptions for text and data mining (TDM). Article 3 provides a mandatory exception allowing research organizations and cultural heritage institutions to engage in TDM without prior authorization, provided they have lawful access to the content. This exception is limited to non-commercial scientific research purposes. Article 4 extends the exception to all users with lawful access, including commercial entities, but introduces an "opt-out" mechanism: the exception applies only when rightsholders have not explicitly reserved their rights [29].

As a result, commercial TDM remains significantly constrained. Rightsholders may contractually exclude their works from TDM by issuing a clear reservation of rights, [30] and, under Article 6(1) of the InfoSoc Directive, users are prohibited from circumventing technological protection measures (TPMs) [31].

In practice, the implementation of this framework encounters two principal barriers. First, database providers often deploy TPMs—such as anti-crawling systems and encryption protocols—to block automated data access. Second, many licensing agreements preemptively prohibit TDM activities, thereby effectively nullifying the applicability of the exception [32]. For instance, platforms such as Elsevier and ScienceDirect explicitly prohibit unauthorized crawling or bulk downloading and require data access exclusively through designated APIs [33]. [34] Although such practices formally comply with the Directive, they reflect rightsholders' strong preference for maintaining granular control over data usage, thereby diminishing the practical utility of Article 4. These dynamics expose a fundamental tension within the DSM framework: while the Directive aims

DOI: 10.54254/2753-7048/2025.HT25556

to facilitate data-driven innovation through limited TDM exemptions, the rights reservation and TPM mechanisms allow rightsholders to reassert control, thereby weakening the intended balance. As such, the current regime fails to fully reconcile the competing interests of data users and copyright owners.

4.2. General clauses and adaptive exceptions in the Japanese framework

In 2018, Japan amended Article 30-4 of its Copyright Act to establish a special exception for information analysis. This provision explicitly permits the reproduction of copyrighted works for the purpose of data analysis without prior authorization. The legal rationale is that such uses are "non-enjoyment-based" and therefore lie outside the core protective scope of copyright. The exception applies regardless of the user's identity or purpose, encompassing a wide array of AI training activities such as natural language processing, image recognition, and other machine learning tasks. Structurally, it adopts a "general clause–illustrative list–residual provision" format, which simultaneously broadens applicability and preserves legal coherence [35].

According to the Japanese Agency for Cultural Affairs, information analysis does not interfere with the normal exploitation of works and should thus be exempt from copyright restrictions [36]. Compared with the European Union's opt-out model, the Japanese framework grants broader rights through a statutory exception, unencumbered by contractual limitations or technological protection measures. Consequently, it exhibits greater adaptability and is more technologically accommodating.

Nevertheless, the Japanese model is not without shortcomings. The scope of permissible use remains somewhat indeterminate, and the absence of a compensation mechanism raises concerns about the fair redistribution of value derived from data use. Although the Japanese approach demonstrates a principled commitment to facilitating technological innovation, it differs fundamentally from the statutory licensing regime proposed in this paper. While Japan's framework emphasizes rights limitation via open-ended exceptions, the model advanced herein advocates for state-authorized usage, paired with post hoc compensation and robust compliance mechanisms. Despite these differences, the semi-open legislative technique embodied in the Japanese model provides a valuable reference point for future copyright reforms in China, particularly in crafting a more balanced and innovation-friendly regulatory environment for generative AI.

5. Designing a statutory licensing framework for text and data mining in China

5.1. The institutional necessity of statutory licensing: insights from history and comparative law

Generative AI models require access to vast and diverse datasets during their training processes. However, under the current legal framework, each stage involving the use of copyright-protected works for data mining carries substantial risk of infringement. In China, the absence of a unified compliance mechanism creates multiple barriers to lawful data utilization. While copyright holders seek to maximize licensing revenues, AI developers strive to acquire extensive content at minimal cost. This inherent conflict of interests, if left unresolved, inevitably increases transaction costs and undermines the fundamental legislative objective of copyright law: the promotion and dissemination of knowledge.

Furthermore, the growing concentration of data resources has given rise to "data monopolization," [37] which restricts equitable access and impedes the free flow of data. For instance, in 2023, Reddit instituted a charge of $12,000 per 50 million API calls, [38] substantially

DOI: 10.54254/2753-7048/2025.HT25556

elevating the cost of data acquisition for AI developers and intensifying competition over training resources. As platform owners tighten control over their data assets, the openness and fairness of the AI ecosystem face escalating threats.

In this context, statutory licensing—historically designed to mediate conflicts between rights holders and users—offers a promising institutional solution to the twin challenges of inefficient authorization and rising data access barriers. By establishing unified authorization standards and centralized compensation mechanisms, statutory licensing can effectively reduce transaction costs and mitigate monopolistic practices. Unlike the TDM exemption model under the EU's DSM Directive—which permits rightsholders to opt out through contractual provisions or technological protection measures, thereby limiting the exemption's scope—a statutory licensing regime can predefine clear conditions for use and standardized compensation rules. This approach circumvents the uncertainty inherent in fair use doctrines and the inefficiencies of individualized negotiations, making it more suitable for the scale and automation demands of AI training.

Although China's Administrative Measures for Generative AI Services mandate that training data be lawfully sourced, the prohibitive cost of securing licenses frequently results in sample bias and algorithmic skew. Implementing a statutory licensing system based on a "use-first, pay-later" principle—facilitated by collective management organizations and emerging technologies such as blockchain—could substantially lower transaction costs, prevent data monopolization, and maintain market openness. Moreover, a mandatory compensation mechanism would harmonize innovation incentives with rights protection, enabling dynamic coordination between data usage and copyright interests.

5.2. Substantive design: legal provisions and institutional mechanisms

5.2.1. A semi-open legislative approach to statutory licensing

From a legislative standpoint, the introduction of a statutory licensing regime for data mining in China would benefit from adopting a semi-open drafting technique that balances legal certainty with technological adaptability. Inspired by Article 30-4 of Japan's Copyright Act, this legislative framework should combine a general clause, an illustrative list, and a residual provision—thereby offering clear guidance for current use cases while preserving sufficient flexibility for future technological advancements.

First, a general clause should be enacted to delineate the fundamental conditions under which data mining activities may proceed without prior authorization but subject to mandatory remuneration. A possible formulation might be: "For purposes of technological research and development, algorithmic optimization, artificial intelligence training, or information analysis, published works may be used without authorization, provided that such use does not interfere with the normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the rightsholder. Fair remuneration shall be paid in accordance with applicable regulations." This clause serves dual functions: it establishes that data mining may qualify for statutory licensing within defined purposes, and it imposes legal boundaries through a "three-part limitation"—the work must be published, the use must not disrupt its normal exploitation, and the use must not cause unreasonable harm to the rightsholder. This structure exemplifies a classic typified general clause balancing access with protection.

Second, the legislation should incorporate an illustrative provision to improve legal clarity and facilitate consistent administrative and judicial application. For example: "The term 'information analysis' as used in the preceding paragraph includes, but is not limited to, text mining, semantic

DOI: 10.54254/2753-7048/2025.HT25556

modeling, image recognition, voice processing, natural language processing, and algorithmic training based on user behavior prediction." By enumerating common data mining scenarios, this provision assists regulators and courts in interpreting the statute consistently, while permitting analogical extension to emerging technologies. This open-ended listing strategy combines legal guidance with adaptive flexibility.

Finally, to accommodate novel forms of data use arising from ongoing technological evolution, a residual provision should be introduced. For instance: "Where a type of use not explicitly listed in this Article is objectively required by the public interest or scientific research, and falls within the functional scope of information analysis, the competent copyright authority under the State Council may formulate supporting administrative regulations or propose amendments to relevant legislation through statutory procedures." This catch-all clause creates a legal mechanism for future applications, maintaining regulatory oversight via state-administered rulemaking. It embodies a dual emphasis on legislative flexibility and administrative accountability.

5.2.2. A new copyright management infrastructure under statutory licensing

Statutory licensing, as a legal authorization independent of rights holders' consent, necessitates robust institutional safeguards to guarantee fair remuneration and effective enforcement. In practice, users may lack sufficient incentives or the capacity to fulfill payment obligations, while rights holders—often hindered by information asymmetries—may struggle to assert their claims for compensation. This dynamic risks undermining the legitimacy and operational effectiveness of the licensing system.

Traditional collective management organizations, which operate based on prior authorization and consent from rights holders, are structurally ill-suited to statutory licensing frameworks where individual consent is not a prerequisite [39]. As Katharina de la Durantaye observes, copyright owners who cannot control the use of their works in AI training also risk losing economic compensation altogether.This challenge highlights the critical need for a novel copyright management mechanism that ensures predictable, enforceable remuneration even absent individual rights holders' direct control [40].

Such an institution would assume responsibility for managing the registration of usage, calculating royalties, identifying rights holders, and distributing compensation. Emerging technologies—such as blockchain and smart contracts—could be leveraged to develop a transparent, auditable system that reliably tracks usage and automates royalty allocation. In cases where ownership is indeterminate or rights holders cannot be identified, a centralized compensation fund should be established to manage and allocate unclaimed revenues. The institutional design of this management body can draw valuable lessons from the European Union's Data Governance Act (DGA), which governs data intermediary platforms. The DGA requires such platforms to remain neutral, undergo official registration and annual reporting, and maintain detailed records of data usage [28]. Although the DGA operates within a voluntary licensing framework, its emphasis on neutrality, transparency, and procedural rigor provides useful guidance for constructing a technically neutral and procedurally robust copyright management institution tailored to China's statutory licensing regime [41].

6. Conclusion

Amid the rapid advancement of generative AI technologies, copyright law must adapt to the evolving landscape of interests and regulatory challenges posed by large-scale data use. Drawing

DOI: 10.54254/2753-7048/2025.HT25556

upon technological imperatives, historical evolution, and comparative legal experiences, this paper contends that when the traditional "authorization–licensing" model proves insufficient to meet the demands for efficient data utilization, statutory licensing emerges as a viable solution to balance creators' rights with industrial innovation. Such a model possesses both normative legitimacy and institutional feasibility. Crucially, statutory licensing does not diminish existing rights but rather reconfigures enforcement mechanisms. By establishing clear rules that ensure lawful use alongside fair compensation, supported by institutional innovation, the overall efficiency of the copyright system can be enhanced. Looking forward, China should develop a statutory licensing framework for text and data mining that aligns with the growth of its data-driven economy. Concurrently, supportive mechanisms—such as rights holder protection and intermediary management systems—must be constructed to foster a synergistic relationship between copyright protection and technological progress.

References

[1] Zittrain, Jonathan, et al. The Files Are in the Computer: Copyright, Memorization, and Generative AI. arXiv, 19 Apr. 2024, https: //arxiv.org/abs/2404.12590. Accessed 29 May 2025.
[2] Vincent, James. "Getty Sues Stability AI for Copying 12M Photos and Imitating Famous Watermark." Ars Technica, 3 Feb. 2023, https: //arstechnica.com/tech-policy/2023/02/getty-sues-stability-ai-for-copying-12m-photos-and-imitating-famous-watermark.
[3] Paul, Kari. "New York Times Sues OpenAI and Microsoft for Copyright Infringement." The Guardian, 27 Dec. 2023, https: //www.theguardian.com/media/2023/dec/27/new-york-times-openai-microsoft-lawsuit.
[4] Zhang, Huibin, and Xiao Qixian. "Constructing Copyright Exemption Rules for Text and Data Mining in the Age of Artificial Intelligence." Science, Technology and Law (Chinese-English Edition), no. 6, 2021, pp. 74–84.
[5] Xu, Xiaoben, and Yang Yinan. "On Fair Use of Copyright in Deep Learning by Artificial Intelligence." SJTU Law Review, no. 3, 2019, pp. 32–42. doi: 10.19375/j.cnki.31-2075/d.2019.03.003.
[6] Wang, Shengjie. "The Legitimate Scope and Infringement Regulation of Text and Data Mining by Generative Artificial Intelligence." Smart Rule of Law, vol. 1, 2024, pp. 213–225. Proceedings of the 2024
[7] World Artificial Intelligence Conference – Legal Forum, edited by Zhongnan University of Economics and Law, 2024. doi: 10.26914/c.cnkihy.2024.009980.
[8] Sun, Ninghui. "Lecture 10 of the 14th National People's Congress Standing Committee Special Lectures: Development of Artificial Intelligence and Intelligent Computing." National People's Congress Official Website, 30 Apr. 2024, http: //www.npc.gov.cn/c2/c30834/202404/t20240430_436915.html.
[9] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.
[10] Awadalla, Anas, et al. "Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens." Advances in Neural Information Processing Systems 37 (2024): 36805-36828.
[11] China Artificial Intelligence Development Report 2024. China Academy of Information and Communications Technology, Dec. 2024, www.caict.ac.cn/kxyi/qwfb/bps/202412/P020241210548865982463.pdf. Accessed 6 Apr. 2025.
[12] de Man, Yvonne, et al. "Opt-in and opt-out consent procedures for the reuse of routinely recorded health data in scientific research and their consequences for consent rate and consent bias: Systematic review." Journal of medical Internet research 25 (2023): e42131.
[13] National People's Congress. Copyright Law of the People's Republic of China, 11th Amendment, 2020. Translated by WIPO, www.wipo.int/wipolex/en/text/603228. Accessed 30 May 2025.
[14] Ma, Zhiguo, and Zhao Long. "The Impact of Text and Data Mining on the Copyright Exception System and the Legal Response." Journal of Northwest Normal University (Social Sciences), vol. 58, no. 4, 2021, pp. 107–115. doi: 10.16783/j.cnki.nwnus.2021.04.012.
[15] World Trade Organization. Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS), 15 Apr. 1994, Article 7. Marrakesh Agreement Establishing the World Trade Organization, Annex 1C. https: //www.wto.org/english/docs_e/legal_e/27-trips.pdf.
[16] Zhan, Ailan, and Tian Yinong. "Copyright Risks in Machine Learning of Generative Artificial Intelligence and Their Resolution Paths." Electronic Intellectual Property, no. 11, 2023, pp. 4–14.

DOI: 10.54254/2753-7048/2025.HT25556

[17] Atkinson, David. "Unfair Learning: GenAI Exceptionalism and Copyright Law." arXiv preprint arXiv: 2504.00955 (2025).
[18] Wang, Guo, and Cao Ziyi. "The Challenges in Determining Fair Use of Copyright in China and the Improvement of Rules: An Analysis Based on 142 Judgments." Journal of Chongqing University of Posts and Telecommunications (Social Sciences Edition), vol. 37, no. 2, 2025, pp. 76–84.
[19] Sherman, Brad, and Lionel Bently. The Making of Modern Intellectual Property Law: The British Experience, 1760 - 1911. Trans. Jin Haijun. Peking University Press, 2006. 12.
[20] Huang, Haifeng. The Expression and Practice of Intellectual Property: A Historical Study of Copyright, Patent, and Trademark. PhD dissertation, Renmin University of China, 2006.
[21] Lipszyc, Delia. Copyright and Neighboring Rights. UNESCO, and China Translation and Publishing Corporation, 2000, p. 18.
[22] Huang, Haifeng. The Discourse and Reality of Intellectual Property: Historical Studies on Copyright, Patent, and Trademark. Huazhong University of Science and Technology Press, 2011.
[23] Li, Chen. Critique of Fundamental Theories of Copyright. Intellectual Property Publishing House.
[24] British Intellectual Property Commission. “Integrating Intellectual Property Rights and Development Policy.” 2002. http: //www.iprcommission.org.
[25] Funatani, Nobuo. Fifty Lectures on Patent Law. Translated by Wei Qixue, Law Press China, 1984, p. 12.
[26] Henn, Harry G. The Compulsory License Provisions of the U.S. Copyright Law. Study No. 5, Copyright Law Revision, prepared for the Subcommittee on Patents, Trademarks, and Copyrights of the Committee on the Judiciary, United States Senate, 86th Cong., 1st sess., Committee Print, U.S. Government Printing Office, 1960.
[27] Mei, Zhenrong, Pei Li, and Wang Hailing. "An Interpretation of the EU Data Governance Act and Its Implications for Data Governance in Smart Libraries in China." Library Journal, no. 4, 2024, pp. 62–68.
[28] Deng, Lingbin. "Legal Regulation of Data Governance: EU Legislation and the 'China Approach'—An Analysis and Reflection Based on the EU Data Governance Act (DGA)." Library Journal, pp. 1–10.
[29] European Union. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market. Recital 14. Official Journal of the European Union, L130, 17 May 2019, pp. 92–125. https: //eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX: 32019L0790.
[30] Directive (EU) 2019/790, Article 7(1) states that “ any contractual provision contrary to the exceptions provided for in Articles 3 […] shall be unenforceable.” See also, Luisa Gatti, ‘The European Solution for Text and Data Mining: A Focus on the Opt-Out System Provided by Article 4 of the DSM Directive’ (2021) 43 European Intellectual Property Review 767.
[31] Gatti (n 61); Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, ‘Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU’ (2019); Nicolas Binctin, ‘TDM: A CHALLENGE FOR ARTIFICIAL INTELLIGENCE’ [2019] Revue In-ternationale du Droit d’Auteur
[32] Geiger, Christophe, Giancarlo Frosio, and Oleksandr Bulayenko. "The exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market-legal aspects." Centre for International Intellectual Property Studies (CEIPI) Research Paper 2018-02 (2018).
[33] "Text and Data Mining Policy." Elsevier, https: //www.elsevier.com/legal/terms-and-conditions/data-mining.
[34] "Terms and Conditions." ScienceDirect, Elsevier https: //www.elsevier.com/en-xm/solutions/sciencedirect/terms-and-conditions.
[35] Li, Chen. "On the Legislative Techniques of 'Fair Use' in the Revision of China’s Copyright Law." Intellectual Property, no. 1, 2013, pp. 12–18.
[36] Isaac M. Reddit. "Reddit Wants to Get Paid for Helping to Teach Big A.I. Systems." The New York Times, 18 Apr. 2023, www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html. Accessed 6 Oct. 2024.
[37] Xie, Qingchuan, and He Tianxiang. "On the Liberalization Path of the Fair Use System of Copyright: Focusing on the Introduction of the 'Intermediate Level' General Clause." Intellectual Property, no. 5, 2019, pp. 58–69.
[38] Agency for Cultural Affairs, Japan. Explanatory Document on the 2018 Amendment to Article 30-4 of the Copyright Act. Agency for Cultural Affairs, 2018.
[39] Tian, Xiaoling. "A Study on the Scope and Related Issues of Collective Management of Copyright: From the Perspective of the Third Amendment to the Copyright Law." Intellectual Property, no. 10, 2015, pp. 89–96.
[40] de la Durantaye, Katharina. "Control and Compensation. A Comparative Analysis of Copyright Exceptions for Training Generative AI." IIC-International Review of Intellectual Property and Competition Law (2025): 1-34.
[41] European Union. Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European Data Governance (Data Governance Act). Official Journal of the European Union, L152, 3 June 2022, pp. 1–44. https: //eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868.