The Copyright Dilemma of Training Data Usage in Generative Artificial Intelligence and Its Resolution: Postprint

Authors: Xie Xingchen, Song Yao, Li Kexin

Date: 2025-06-20T00:00:00+00:00

Abstract

As training data for generative artificial intelligence constitutes a foundational resource for technological innovation, its compliant utilization carries strategic significance for promoting algorithm optimization and industrial iteration. Nevertheless, the rules concerning licensed use, fair use, and statutory licensing under the traditional copyright framework have become severely inadequate; the massive data requirements of generative artificial intelligence have created a conflict with the existing copyright system, which has evolved into legal shackles that constrain innovation in the artificial intelligence industry. Through normative analysis and comparative study, this article elaborates in detail on the copyright predicament and its underlying causes regarding the use of training data for generative artificial intelligence. Based on a critical examination of institutional practices in the United States, Europe, and Japan, it proposes a three-pronged path for constructing a copyright exception system for generative artificial intelligence training data in China: first, reconstructing fair use rules by incorporating "information-analysis use" into the exemption scope and establishing a "no market conflict" criterion; second, innovating a quasi-statutory licensing system to establish a flexible authorization pathway via a "notice-and-objection exclusion" mechanism; and third, exploring the copyright collective management organization approach to construct a large-scale authorization framework featuring "default licensing plus precise profit distribution." This serves to mitigate the conflict between rights protection and industrial development, preventing the system from stifling innovation while also forestalling innovation from eroding rights.

Full Text

Abstract

Training data for generative artificial intelligence (GenAI) serves as a foundational resource for technological innovation, and its compliant utilization holds strategic significance for driving algorithmic optimization and industrial iteration. However, traditional copyright frameworks—centered on authorized use, fair use, and statutory licensing—are increasingly inadequate. The massive data demands of GenAI have created conflicts with existing copyright systems, evolving into legal constraints that hinder AI-driven innovation. Through normative analysis and comparative study, this article elaborates on the copyright dilemmas and underlying causes associated with GenAI training data use. Based on a critical examination of institutional practices in the United States, Europe, and Japan, we propose a tripartite approach to constructing a copyright exception system for GenAI training data in China: First, restructure fair use rules by incorporating "information-analytical use" into exempted categories and establishing a "no market conflict" criterion; second, innovate a quasi-statutory licensing system through a "public notice + objection exclusion" mechanism to create a flexible authorization pathway; third, explore collective management organization models to build a scalable authorization system based on "default licensing + precise profit-sharing." These proposals aim to reconcile the tension between rights protection and industrial development, preventing institutional suppression of innovation while simultaneously guarding against innovation eroding rights.

Keywords: Generative artificial intelligence; Training data; Authorized use; Fair use; Statutory license

Introduction

Accelerating the development of next-generation artificial intelligence represents a strategic imperative for China to seize opportunities in the new round of technological revolution and industrial transformation. General Secretary Xi Jinping has emphasized the importance of the healthy development and regulated application of GenAI. Policy documents such as the *Interim Measures for Generative AI Services* and the *Measures for the Identification of AI-Generated Synthetic Content* demonstrate the state' s high-level attention to GenAI development. As GenAI technologies like ChatGPT and DeepSeek achieve breakthrough advances, their training processes exhibit growing dependence on large-scale collections of copyrighted works, triggering copyright compliance disputes that have become legal obstacles constraining industry development. The tension between technological iteration and legal regulation continues to intensify. Whether from the perspective of authorized use, fair use, or statutory licensing, the current copyright system struggles to provide clear compliance pathways for GenAI training data use. This unresolved question demands a copyright exception system for resolution. While existing research offers multidimensional

perspectives on addressing these challenges [6-10] and discusses potential regulatory directions drawing on foreign experiences [11-14], this article systematically examines the copyright dilemmas of GenAI training data use and proposes solutions grounded in China' s national context.

1. Copyright Dilemmas in GenAI Training Data Use

1.1 Authorized Use Dilemmas

China' s Copyright Law stipulates that using copyrighted works requires prior authorization and payment obligations. However, the traditional authorization framework based on case-by-case negotiation cannot effectively adapt to GenAI' s characteristics. First, the data scale required for GenAI training far exceeds the capacity of conventional authorization mechanisms. Obtaining authorization for each work individually would be prohibitively time-consuming. Second, the financial costs are substantial. Data demanders must not only accurately identify rights holders—already a costly endeavor—but also negotiate authorization scope repeatedly. When ownership is unclear, costs escalate dramatically. Third, rights traceability creates legal chain-of-title fractures. Data sourced from web scraping often lacks original rights holder information. Even when identified, the sheer number of rights holders drastically reduces authorization efficiency. In film production, for instance, all creative contributors—screenwriters, directors, cinematographers, composers—may become legal authors [23]. This dispersion of authorization subjects across temporal and geographic dimensions fundamentally conflicts with GenAI' s need for massive datasets.

1.2 Fair Use Rule Application Dilemmas

China' s current legal framework lacks specific guidance on fair use during data training, forcing reliance on traditional copyright law. Debates persist over whether training constitutes reproduction or adaptation. Reproduction right controversies focus on the digitization of works during data input. While China' s Copyright Law Article 10 includes temporary copies within the reproduction right scope, some scholars argue that GenAI' s transient storage of data for technical purposes constitutes a non-substantial, technical intermediate process that doesn' t infringe reproduction rights [16]. Adaptation right disputes center on whether model training constitutes modification. Proponents argue that if output content retains the core expression of original works, it may infringe adaptation rights [15], while opponents note that GenAI doesn' t subjectively modify originals, merely analyzing relationships between works to reconstruct patterns [18]. The U.S. Copyright Office' s *Copyright and Artificial Intelligence* report suggests a middle ground: imitation of artistic style may not infringe copyright, but identifiable original work fragments in output could constitute infringement [19].

These controversies create a compliance vacuum. GenAI developers cannot clearly identify which rights to request or build compliant authorization path-

ways, rendering the prior authorization principle practically inapplicable. Even attempting to apply fair use through legal interpretation proves strained. China' s fair use provisions strictly limit eligible subjects to natural persons for personal learning, research, or appreciation—purposes distinct from GenAI systems' commercial training objectives. Although algorithm training could qualify as research, the law restricts eligible entities to public educational and research institutions, excluding commercial developers. The EU' s Digital Single Market Copyright Directive, which allows text and data mining exceptions for research institutions and cultural heritage organizations, similarly cannot be extended to GenAI [24].

1.3 Statutory Licensing Dilemmas

Statutory licensing systems, designed to correct market failure, face structural incompatibility with GenAI training data applications. Proponents emphasize efficiency gains and reasonable compensation for copyright holders [26-28], yet fail to systematically address the misalignment between GenAI' s technical characteristics and statutory licensing architecture. Drawing on U.S. copyright law evolution, statutory licensing serves as a transitional buffer during technological upheaval, operating within—rather than disrupting—market autonomy frameworks [29]. However, this design may exacerbate enforcement challenges given the concealed nature of data use, making rights holder evidence collection difficult.

Pricing mechanism rigidity presents another obstacle. Effective statutory licensing for GenAI training requires dynamic fee adjustment systems balancing multiple dimensions: model commercial value, operational costs, and market adaptation. China' s rapid technological iteration continuously disrupts data market supply-demand relationships and pricing structures. Current statutory licensing frameworks lack the flexibility to accommodate these dynamics, creating legislative and practical implementation challenges for GenAI training data applications that require systematic analysis of billions of works.

1.4 Judicial Practice Dilemmas

Judicial determinations of fair use for GenAI training data remain challenging. While litigation has emerged, existing cases either fail to address core controversies or involve special circumstances that limit their precedential value. The "Ultraman" case is illustrative: the court found infringement in the generation phase but rejected the plaintiff' s demand to delete training data, reasoning that copyright doesn't grant exclusive rights over model training post-use [25]. While guiding for generation-phase regulation, the decision didn' t address training-phase infringement or fair use application.

The "voice personality rights" case similarly sidestepped copyright issues. The court creatively determined that unauthorized use of a voice actor' s voice for model training constituted personality rights infringement, noting that copy-

right doesn' t include exclusive rights to prevent others from training models to generate corresponding outputs [25]. The case' s focus on personal attributes rather than copyrighted works limits its 参考价值. These judicial approaches reflect conservative attitudes toward new technology' s copyright impacts, leaving GenAI developers uncertain about training data legality.

2. International Copyright Exception Systems for GenAI Training Data

2.1 United States: Restrictive Fair Use Determination

U.S. fair use doctrine, anchored in the four-factor test, provides the traditional benchmark for determining fair use. In the Google Books case, courts held that mass digitization for search and snippet creation constituted highly transformative fair use [32]. However, recent cases show a trend toward stricter interpretation. The Thomson Reuters case, concerning legal summary training for competitive products, saw courts initially deny fair use claims [33]. This judicial conservatism reflects cautious U.S. judicial responses to copyright challenges posed by new technologies.

2.2 Japan: Information Analysis Fair Use Framework

Japan' s 2018 copyright amendment established an innovative framework for AI training data. Article 30-4 of Japan' s Copyright Law distinguishes between "enjoyment use" (appreciating works' ideas and aesthetics) and "non-enjoyment use" (technical functions without accessing spiritual value). Information analysis —such as word frequency statistics, waveform analysis, or pixel parsing—qualifies as non-enjoyment use because it doesn' t provide emotional experience or affect copyright holders' market interests [12]. The law also exempts incidental copying during computer information analysis and minimal use when providing analysis results, addressing gaps in Article 30-4 [12].

However, the framework has limitations. Ambiguity in quantifying GenAI output content may lead to inconsistent judicial applications. Some Japanese commentators argue the rules are overly permissive, inadequately protecting copyright holders [38].

2.3 European Union: Text and Data Mining Exceptions

The EU' s Directive 2019/790 establishes a two-tier TDM exception system. Article 3 creates a mandatory exception for research institutions and cultural heritage organizations, allowing TDM on lawfully accessed works while prohibiting contractual opt-outs. Article 4 provides an optional commercial exception permitting rights holders to reserve their rights through technical measures or contractual declarations [13]. This 分层设计 reflects legislators' value choices between innovation incentives and private rights protection.

The EU system requires lawful data sources and prohibits uses that conflict with normal work exploitation. While providing legal certainty for data mining, it allows member states flexibility to expand exceptions based on digital industry needs [35]. However, the fragmented implementation across member states and rights holder opt-out mechanisms for commercial use create uncertainty for global GenAI model training.

3. Constructing China' s Copyright Exception System for GenAI Training Data

3.1 Restructuring Fair Use Rules

Drawing on Japan' s experience, China should incorporate "information-analytical use" into its fair use system through legislative interpretation or judicial guidance. The core transformation involves distinguishing between expressive and functional uses. Information analysis focuses on extracting linguistic rules, image structures, and information features without accessing works' substantive value, typically not affecting original markets. Establishing "no market conflict" as a central criterion aligns with international law foundations and China' s development needs.

Fair use can provide a safety net for non-commercial, non-competitive GenAI research, preventing excessive restriction of technological exploration. However, deeper issues of data utilization efficiency and interest balancing require complementary mechanisms through quasi-statutory licensing and collective management organizations (CMOs).

3.2 Innovating Quasi-Statutory Licensing

Rigid statutory licensing requires explicit legal authorization and denies copyright holders opt-out rights, lacking necessary flexibility for GenAI training data scenarios. A quasi-statutory licensing model offers greater adaptability. Building upon the framework of China' s *Regulation on the Protection of Information Network Transmission Rights*, the core lies in establishing a "public notice + objection exclusion" mechanism.

Developers should fulfill mandatory disclosure obligations by publishing lists of works intended for training data on designated platforms, specifying information analysis purposes, compensation standards, and other key terms. A notice period of no less than 30 days should allow copyright holders to exercise objection rights. Works explicitly opposed during the notice period must be removed by developers. For remaining works, developers may proceed with training after paying 阶梯化 compensation based on industry conventions.

This architecture reduces rights confirmation costs through notice procedures while preserving copyright holder control via opt-out rights. It maintains creation incentives while promoting technological development. Implementation

requires detailed supporting rules through judicial interpretation or administrative regulations, including dynamic filtering mechanisms for opted-out works and mandatory record-keeping for compliance review.

3.3 Exploring Collective Management Organization Pathways

Compared to traditional uses, GenAI training demands for work quantity, variety, and usage frequency increase exponentially. Individual authorization models face prohibitive transaction costs and efficiency bottlenecks. While CMOs traditionally addressed mass licensing in music and publishing, their functions and market capabilities require restructuring to handle GenAI training scenarios.

A "default licensing + precise profit-sharing" system could provide a key pathway for data compliance and industrial development. Under this model, all CMO-managed works would permit non-expressive data use by default, with standardized licensing fees and usage scopes established through collective bargaining. Rights holders could declare prohibitions on training use, requiring real-time filtering systems to ensure timely removal. This approach avoids the unpredictability of U.S. case-by-case fair use determinations, overcomes EU' s binary system fragmentation, and provides a normative model for China' s participation in global AI governance.

4. Conclusion

Copyright compliance for GenAI training data represents a value rebalancing process between creative ecosystems and technological innovation in the digital age. Deconstructing traditional copyright systems' responses to technological change reveals three layers of dilemmas caused by legal lag. Through critical examination of foreign experiences, this article provides new solutions for constructing a copyright exception system adapted to China' s national conditions.

As China' s GenAI industry reaches a critical development juncture, timely legal adjustment and improvement will determine China' s position in global technological competition. The proposed tripartite framework—restructured fair use, quasi-statutory licensing, and collective management—seeks equilibrium between protecting rights, promoting industry, and safeguarding diverse interests. This balanced approach charts a course toward harmonious coexistence between technology and law, ensuring China' s new-generation AI develops robustly within legal bounds.

References

[1] Research on the Modernization of Industrial Intellectual Property Risk Governance Under the Overall National Security Concept (National Social Science Fund Major Project: 21&ZD204)

[2] Generative AI Training Data Copyright Dilemma and Resolution. *Publishing and Distribution Research*, 2024(12):91-97.

[3] Institutional Challenges and Solutions for Copyright Legitimacy of AI-Generated Content. *Journal of Northwest University of Political Science and Law*, 2024(3):18-31.

[4] Non-Work Use in GenAI Data Training: Legitimacy Justification. *Library and Information Knowledge*, 2024(3):67-78.

[5] Copyright Law Response to GenAI Training Data. *Journal of Intelligence*, 2025(1):78-88.

[6] Does Fair Use Apply to Large Model Data Training? *East China University of Political Science and Law Journal*, 2024(4):20-33.

[7] Copyright Infringement by GenAI. *Electronic Intellectual Property*, 2024(11):37-58.

[8] Copyrightability Issues of AI Creations. *Intellectual Property*, 2020(3):653-673.

[9] Qualification of AI-Generated Content Under Copyright Law. *Journal of Northwest University of Political Science and Law*, 2017(5):148-155.

[10] Institutional Challenges and Solutions for Copyright Legitimacy of AI-Generated Content. *Journal of Northwest University of Political Science and Law*, 2024(3):18-31.

[11] Copyright Infringement Risk Regulation for AI Training Data: Local Dilemmas and Solutions. *Publishing and Distribution Research*, 2025(1):82-99,150-151.

[12] Japan' s GenAI Training Data Fair Use Rules. *Library Forum*, 1-9 [2025-03-06]. https://link.cnki.net/urlid/44.1306.g2.20250224.1351.004.

[13] Copyright Law Response to GenAI Training Data: EU Copyright Exception Rules and Implications for China. *Library Forum*, 1-11 [2025-02-06]. https://link.cnki.net/urlid/44.1306.G2.20250115.1117.002.

[14] Copyrightability of AI Creations: Japanese Experience and Chinese Path. *Contemporary Japanese Economy*, 2025(1):81-94.

[15] Fair Use Rules for GenAI Data Training. *Journal of East China University of Political Science and Law*, 2024(4):20-35.

[16] Copyright Infringement Risks and Legal Responses in GenAI Training. *Journal of Xiangtan University (Philosophy and Social Sciences Edition)*, 2024(5):78-86.

[17] Copyright Risks and Mitigation in AI Creation Data Acquisition and Use. *Chongqing Social Sciences*, 2022(4):128-140.

[18] Copyright Rules for Machine Learning: Historical Insights and Contemporary Solutions. *Intellectual Property*, 2023(6):97-113.

[19] United States Copyright Office. *Copyright and Artificial Intelligence Part 1: Digital Replicas* [EB/OL]. [2025-02-06]. https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Digital-Replicas-Report.pdf.

[20] Principle and Construction of GenAI Copyright Compensation System. *Publishing and Printing*, 2025(1):37-47.

[21] Challenges and Responses of AI to Copyright Limitation and Exception Rules. *Intellectual Property*, 2022(11):152-162.

[22] Relationship Between Copyright System Evolution and Cultural Industry Transformation: Analysis of Film Copyright System. *Chongqing Social Sciences*, 2022(11):127-139.

[23] Regulation of GenAI Training Copyright Disputes. *China Copyright*, 2024(8):63-71.

[24] First-instance Judgment in National First GenAI Voice Infringement Case [EB/OL]. [2025-02-06]. https://www.chinanews.com.cn/sh/2024/04-25/10205621.shtml.

[25] Copyright Infringement Issues in Machine Learning and Solutions. *Intellectual Property*, 2019(2):68-79.

[26] Challenges and Responses of Machine Learning to Copyright Fair Use System. *Intellectual Property*, 2020(10):13-25.

[27] Legal Qualification and Protection Path of AI Compilation Use from Perspective of Interest Balance. *Publishing and Distribution Research*, 2020(11):72-79.

[28] Evolution of U.S. Fair Use System Since Campbell Case. *China Copyright*, 2016(12):82-90.

[29] New Developments in U.S. Copyright Law: Commentary on Authors Guild v. Google. *China Copyright*, 2014(1):58-60.

[30] "Unfair Competition" in Data Training–Thomson Reuters v. Anthropic Intelligence [EB/OL]. [2025-03-01]. https://mp.weixin.qq.com/s/iQo_{XaP5IwHK1OXm0ae4jw}.

[31] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994).

[32] Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).

[33] Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc., No. 1:20-cv-613-SB (D. Del. 2023).

[34] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 [EB/OL]. [2025-02-06]. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0790.

[35] World First Dataset Creation Copyright Infringement Case: Text and Data Mining? [EB/OL]. [2025-02-06]. https://mp.weixin.qq.com/s/uVczdzYH3HeKfStd1442Ow.

[36] Society 5.0 [EB/OL]. [2025-02-06]. https://www8.cao.go.jp/cstp/society5_0/.

[37] Japan's Flexible Fair Use Clause and Its Implications. *Intellectual Property*, 2022(1):112-130.

[38] "Too Permissive"? Japanese Domestic Views on GenAI Regulations [EB/OL]. [2025-02-06]. https://ascii.jp/elem/000/004/122/4122855/.

[39] Economic Analysis of Machine Data Use as Copyright Fair Use. *Intellectual Property*, 2024(3):107-126.

[40] Copyright Law Response to Non-Expressive Use of Works in Digital Context. *Intellectual Property*, 2024(9):110-126.

Note: Figure translations are in progress. See original paper for figures.