Vol.6 Iss.2 2025
Modern Law Research
https://doi.org/10.37420/j.mlr.2025.013

Research on Copyright of Generative Artificial Intelligence for Promoting Data Open Utilization

Weilin Wang*
University of International Relations, Law, Beijing, China
***Corresponding author, E-mail: wwl002713@163.com**

Abstract

This article, through understanding the research status and frontier dynamics in the fields of artificial intelligence copyright and data openness, conducts a detailed analysis of typical cases such as the "Tremblay v. OpenAI class action" and the "Getty Images Corporation v. Stability AI Corporation" cases, revealing from a practical level the contradictions and actual issues between data openness and copyright protection. Meanwhile, by comparing the differences in data governance paths and legislation between the European Union and the United States, it summarizes experiences and lessons, exploring development paths and institutional models suitable for China based on foreign advanced experiences, fully combining with China's national conditions and industrial development needs, and proposes suggestions for improving copyright to promote the openness of generative artificial intelligence data, providing a unique perspective and ideas for the healthy development of China's generative artificial intelligence industry.

Keywords

Generative Artificial Intelligence, Copyright, Training Data

1 Preamble

1.1 Research Background

In the current digital era, generative artificial intelligence technology has achieved groundbreaking progress, and the emergence of models such as ChatGPT and DeepSeek has brought unprecedented opportunities for social development. These technologies can generate realistic text, images, audio, and other content, demonstrating enormous application potential in numerous fields, from content creation assistance to intelligent customer service, from medical diagnosis to financial risk assessment, playing an increasingly important role.

With the vigorous development of generative artificial intelligence, the legal issues of data in the training phase have gradually become key factors constraining the further development and application of technology. Data, as the core element of generative artificial intelligence, is akin to blood to the human body, and its acquisition, processing, and usage methods directly relate to the performance and output quality of artificial intelligence models. In the data-driven development process of artificial intelligence, the contradiction between data openness and copyright protection is increasingly prominent. On one hand, the training of artificial intelligence models requires massive data support, and sufficient openness of data can promote continuous optimization and innovation of models, driving rapid technological development; on the other hand, copyright protection

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

ensures the lawful rights and interests of creators, motivating them to continuously create high-quality works. If the balance between the two is disrupted, it will not only harm the enthusiasm of creators and suppress the source of cultural innovation but also potentially trigger numerous legal disputes, hindering the healthy development of the artificial intelligence industry, thereby affecting the fairness and justice and economic order of the entire society.

1.2 Research Purpose and Significance

This research aims to deeply explore how to reasonably protect generative artificial intelligence copyright while promoting data openness, construct a balanced data governance system and framework of the law, achieve a positive interaction between technological innovation and intellectual property protection, and promote the sustainable development of the artificial intelligence industry.

Researching this issue theoretically contributes to enriching and perfecting the intellectual property theoretical system in the field of artificial intelligence. In-depth research on the relationship between data and generative artificial intelligence copyright can fill some gaps in existing theories in this emerging field, providing a theoretical foundation for further exploration of legal issues in the development of artificial intelligence technology. By analyzing the data governance models and legislative experiences of different countries and regions, it provides a reference for constructing a universal and forward-looking artificial intelligence copyright theory.

In practice, it can provide clear legal guidance and practical guidance for the healthy development of the artificial intelligence industry. By clarifying data usage rules, copyright vesting, and standards for determining infringement, it helps enterprises avoid legal risks in the development and application of generative artificial intelligence technology, guaranteeing the lawful rights and interests of data providers, copyright owners, and technology developers. It promotes the reasonable flow and sharing of data, improves the utilization efficiency of data resources, and promotes the widespread application of artificial intelligence technology in various industries, injecting new impetus into social and economic development. The research results can also provide a decision-making basis for government departments to formulate relevant policies and statutes, strengthen the supervisory and specification of the artificial intelligence industry, and create a good innovation environment and market order.

1.3 Current Research Status at Home and Abroad

Abroad, research on artificial intelligence copyright and data openness and protection started earlier and has yielded fruitful results.Many scholars and research institutions from the dimensions of law, ethics, and others deeply explore the copyright vesting issue of content generated by artificial intelligence. The European Union actively organizes interdisciplinary teams, through project funding and other means, to deeply study the impact of artificial intelligence technology on the traditional copyright system and attempts to propose adaptive framework of the law suggestions. The European Union General Data Protection Regulation (GDPR) has made strict specifications on data protection, emphasizing the Rights of the data subject and specifying the behavior of data processors. The research in the United States starts from different creation modes, analyzing the Standards for qualifying originality of works and the principles of copyright

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

vesting. In terms of data openness, the United States government has established open data platforms such as Data.gov, formulated relevant policy Standards, and promoted the opening of government data to the public to promote innovation and economic development.

In recent years, domestic research in related fields has also shown a vigorous development trend. Scholars closely combine the actual development of China's artificial intelligence industry, focusing on the legal qualification of artificial intelligence generated content under the domestic law framework. With the rapid rise of the domestic artificial intelligence industry and the continuous emergence of related cases, research pays more attention to analyzing actual cases, exploring how to interpret and handle new issues arising from artificial intelligence creation in the existing copyright legal system, such as the copyrightability of artificial intelligence created works, Rights vesting, and Standards for qualifying Infringement.The government of our country actively advocates and promotes the opening of government data, and various levels of government departments have successively established data open platforms and introduced a series of policy measures to promote data openness. The domestic academic community is also strengthening theoretical research on data openness, exploring models of data openness and issues such as data security and privacy protection. In terms of data protection, the continuously improving relevant laws and regulations reflect our country's emphasis and efforts in data protection.

2 Generative Artificial Intelligence Data Training and Related Copyright Case Analysis

2.1 Generative Artificial Intelligence and Data Training

Generative artificial intelligence is a type of artificial intelligence technology capable of autonomously generating new content by learning patterns and rules from existing data. Its core technical principles include deep learning and neural networks, through constructing complex model structures to train on massive data, thereby achieving the learning and understanding of data characteristics. In text generation, the model will learn the usage of vocabulary, grammatical structure, and semantic information, thus being able to generate coherent and logically reasonable text content based on a given theme or prompt. In image generation, the model can learn features such as color, shape, and texture from a large amount of image data, thereby generating images with similar features or entirely new creative images.

Generative artificial intelligence's demand for data exhibits characteristics of large-scale, diversity, and high quality. Large-scale data volume helps the model fully learn various patterns and rules, improving the accuracy and richness of generated content. OpenAI's model uses a large amount of internet text data for training, with its parameter scale reaching hundreds of billions, thus being able to generate high-quality text content. Diverse data can expose the model to different types, styles, and themes of information, avoiding the singularity and bias of generated content. High-quality data requires data to be accurate, complete, and in compliance with statutes and ethical specifications, to ensure the model learns correct knowledge and patterns, avoiding generating incorrect or harmful content.

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

The training of artificial intelligence systems relies on a large amount of data, with various acquisition methods, including public datasets, data obtained through web crawlers, and user-provided data.However, the source and usage of data often involve complex legal issues. Data obtained by web crawlers may involve capturing content from others' websites, and whether this constitutes infringing is a matter of CONTROVERSY; when using public datasets, there may also be ambiguities regarding the scope of Authorization and Limitations on use.

2.2 The Importance of Data

2.2.1 Data is the Creative Source of Artificial Intelligence Outputs

Data, as the creative source of artificial intelligence outputs, directly determines the quality and innovativeness of the outputs through its quality, quantity, and diversity. Rich and diverse training data enables artificial intelligence models to learn more patterns and rules, thereby generating content with greater creativity and value. An artificial intelligence painting model with a vast amount of high-quality image data can generate more realistic and unique painting works, exhibiting higher standards in color usage and composition design; conversely, systems with scarce or singular data often produce works of lower quality and lack creativity, such as models trained with only a small number of specific style images, which may generate paintings with a singular style and monotonous content.

2.2.2 Data Influences the Copyrightability Judgment of Artificial Intelligence Outputs

In judging whether artificial intelligence outputs possess copyrightability, the legality and Compliance of data play a crucial role. If the use of data violates relevant Statutes or infringes others' Rights, then the content generated based on such data may be qualified as lacking copyrightability.Using unauthorized copyrighted works as training data, the generated artificial intelligence outputs may be considered infringing works, unable to obtain copyright protection. In contrast, legally obtained and used data helps support the originality and inventiveness of artificial intelligence outputs, enhancing their copyright qualification. If the data source is legal and reasonably processed, the generated content to some extent reflects the innovative capability of the artificial intelligence model, such as unique news commentary generated from news data based on Due Authorization, which may possess copyrightability.

2.2.3 Data determines the value and market competitiveness of artificial intelligence outputs

High-quality data enables artificial intelligence outputs to better meet market demands, thereby possessing higher value and market competitiveness. In commercial applications, artificial intelligence works generated based on quality data can attract more users and Clients, bringing greater economic benefits to developers and users. High-quality articles or stories generated by artificial intelligence can attract more readers, enhance the traffic and influence of media platforms, and thereby increase advertising Revenue and other commercial proceeds. In the design field, unique and exquisite artificial intelligence-generated images can meet the needs of corporate brand promotion and product design, obtaining higher commercial value and enhancing the market competitiveness of products.

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

2.3 Case Analysis Related to Generative Artificial Intelligence Copyright

2.3.1 "Tremblay v. OpenAI Class Action" case

On June 28, 2023, renowned authors Mona Awad and Paul Tremblay jointly filed a lawsuit against OpenAI.OpenAI in artificial intelligence system training, was accused of extensively ingesting Plaintiff's copyright-protected works, such as using datasets like BookCorpus, which contain numerous unpublished but copyrighted works, for training its GPT series models. Plaintiff discovered that when prompting ChatGPT with the title of their work, it could provide detailed introductions and summaries, thus believing that their novel was used as training material. Plaintiff asserts that the large language model used by ChatGPT infringed upon the reproduction rights and adaptation rights of the work's authors, as the model relies on the existence of a large amount of work content, constituting derivative or derivative Infringement.

2.3.2 "Getty Images Corporation v. Stability AI, Inc." case

In February 2023, the internationally renowned photo library operator Getty Images Corporation sued Stability AI, Inc. Defendant was accused of unauthorized reproduction of more than 12 million still-copyrighted photographic works from Getty Images Corporation's library, used for training its "Stable Diffusion" artificial intelligence software system. This software system, by utilizing these photographic works and accompanying brief descriptions, assists users in machine learning, establishing associations between images and text, achieving subsequent adaptation use, and profiting from Charges, forming a direct competitive relationship with Plaintiff. Furthermore, Getty Images Corporation also accused Defendant of infringing its Trade Marks rights and engaging in Unfair Competition.In the above two issues, OpenAI and other artificial intelligence development enterprises are accused of using a large number of copyright-protected works for artificial intelligence system training without obtaining Licenses and making Payments. This directly reflects the contradiction between the demand for data in the development process of generative artificial intelligence and copyright protection, highlighting the importance of researching how to protect the rights and interests of copyright owners while meeting the data needs of artificial intelligence, in order to achieve a balance between data openness and protection.

Enterprises, in pursuit of technological progress and commercial interests, often neglect the rights and interests of copyright owners, and the openness and utilization of data lack reasonable specification and supervisory. For copyright owners, safeguarding rights faces many challenges, such as the difficulty of providing evidence of Infringement and the uncertainty of Statutes application. These cases enlighten us on the need to construct a more complete data governance system and legal system, clarify data usage rules, strengthen the protection of copyright, and promote the balance between data openness and copyright protection, to achieve the healthy and sustainable development of the artificial intelligence industry.

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

3 Generative Artificial Intelligence Facing Data Copyright Challenges

3.1 Dispute Arising from Legitimacy of Data Sources

In the current digital era, one of the primary challenges faced by Generative Artificial Intelligence Copyright Protection is the issue of the legitimacy of data sources. The data sources used during the training of artificial intelligence systems are extremely broad, covering various types such as proprietary data, open-source datasets, externally sourced data, data collected through automated means, and synthetic data. However, these data may contain Personal Information processed without the explicit Authorization or beyond the scope of Authorization of the data owners, violating relevant privacy protection regulations; they may also involve Violation of open-source dataset License Agreements, and illegal acquisition of computer Information system data and other unlawful acts. The existence of these situations makes the legitimacy of data sources elusive, and the issue of Copyright Ownership of Generated Content generated based on these Data Training becomes exceptionally complex. When works are generated by a model trained with data obtained through illegal means, should the Copyright belong to the original data owners or the developers of the model? This issue presents significant legal Dispute. This uncertainty not only poses challenges to the Vesting of Copyright but also easily triggers a series of Copyright disputes, bringing legal risks and moral dilemmas to the Interested Party.

Data Openness requires data to be widely accessible and usable, but under the Copyright framework, the use of data requires Authorization. Currently, the process of obtaining Authorization is cumbersome and costly, especially for the use of large-scale data.Data providers often worry that data may be misused after Authorization, leading to damage to their own rights and interests, and thus are cautious during Authorization, which seriously hinders the smooth flow of data and Limits the development of Generative Artificial Intelligence.

3.2 Unclear scope of Fair Use

The traditional Copyright Fair Use system is difficult to directly apply when facing the emerging technology of Generative Artificial Intelligence. The methods and scale of artificial intelligence technology in processing and utilizing data have essential differences from traditional human creative activities. This difference leads to the existing Fair Use Standards being unable to provide clear and accurate judgments when assessing the behavior of artificial intelligence. Especially when it involves artificial intelligence learning from a large number of copyright-protected works and generating new content based on this foundation, whether it can be classified as Fair Use is currently a subject of widespread CONTROVERSY in the legal and academic fields. This uncertainty brings considerable legal risks to data users who wish to innovate and develop using data, as they may inadvertently infringe on others' copyrights, thereby facing legal Litigation and Indemnification Liability.

Unclear scope of Fair Use easily leads to improper public data scraping during Generative Artificial Intelligence Data Training.Many training datasets are obtained from public channels, but these data may contain materials without appropriate Licenses or Authorizations, which not only may constitute a Violation of the

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

Rights of data providers but also easily trigger disputes and legal disputes regarding the Fair Use of public data. Significant differences exist in the legal positions of different Countries and regions concerning public data. The European Union has implemented strict protection measures for public Personal Information, aiming to safeguard Individuals' privacy Rights, while the United States considers certain publicly available Personal Data as exceptions to Personal Data protection, which brings different legal requirements and potential risks for data scraping and use. Enterprises, when conducting data scraping, if failing to comply with relevant Statutes and regulations, may lead to an infringement of data providers' Copyright.

3.3 The dilemma of balancing Personal Data Rights protection and artificial intelligence Copyright

In the training of Generative Artificial Intelligence, there exists a dilemma in balancing Personal Data Rights protection and Copyright Protection. A large amount of Personal Data is collected and used, and if protection measures are inadequate, it may lead to the leakage of personal privacy, misuse of Personal Information, and other issues, while also potentially affecting the control of the data subject over their Personal Data in artificial intelligence creation. When users request an Interpretation of automated decisions, it may involve reviewing training data, which conflicts with the privacy Rights of the data source subject, making it difficult to find a balance between protecting Personal Data Rights and ensuring the Fair Use of artificial intelligence Copyright.

4 Analysis and Reference of Foreign Generative Artificial Intelligence Copyright Data Governance Pathways

4.1 Analysis and Reference of European Union Classification, Grading and Subject-based Concept

European Union implements a classification, grading, and subject-based strategy in Generative Artificial Intelligence Data Governance, focusing on the governance of High-risk AI Systems. This concept emphasizes comprehensive specification of Datasets and their collection and processing activities, with Transparency, Purpose Principle, Proportionality Principle, and Anti-discrimination as core governance dimensions. According to the European Union's Artificial Intelligence Act, providers of High-risk AI Systems must completely record the entire data processing process, including the initial purpose of data collection, preparation operations such as Data Annotation, Marking and Cleaning, and the Assessment of the availability, quantity, and suitability of Datasets. Foundation Model providers must strictly screen governance data and meticulously verify data sources; Generative AI Providers must also disclose the usage status of Copyright content. In terms of Datasets quality control, Training Data Management revolves around the expected purpose of AI systems, adhering to core governance principles to ensure data relevance and representativeness, complete error review, and achieve integrity in specific scenarios.

This concept of European Union provides a comprehensive and detailed Data Governance approach for our Country. Our Country can formulate data classification and grading Standards that align with national

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

conditions, clarifying processing requirements and supervisory focus for different levels of data. For critical data involving national security and personal sensitive Information, more stringent management measures should be implemented.In terms of main responsibility, further refine the respective duties of data providers, processors, and users, strengthen the supervision of the entire data processing process, ensure that data processing activities are legal, compliant, and meet ethical standards. Meanwhile, attention should also be paid to improving the Transparency of data processing, by establishing an Information Disclosure mechanism, allowing data subjects to clearly understand the usage of data, and enhancing public trust in Generative Artificial Intelligence technology.

4.2 Analysis and Reference of Transparency of Training Data in the European Union

European Union and Member States regard training data Transparency as a core focus of Data Governance. The European Union Commission's "Guidelines on Artificial Intelligence and Data Protection" requires a strict Assessment of the nature and quantity of data during the training phase, reduction of redundant data, and gradual expansion of the training set scale; simultaneously exploring Automatic Forgetting Algorithms and anonymization Synthetic Data technology to reduce potential risks to personal Rights in data processing. In the practice of Member States, the French CNIL (Commission Nationale de l'Informatique et des Libertés) conducts supervision around data sources, legal basis, sensitive data processing, Data Minimization, and other dimensions, providing specific guidance on the construction of scientific research databases, application of the Purpose Principle, and division of main responsibility; Italy focuses on information transparency, legal basis compliance, data accuracy, and protection of minors; the United Kingdom also advocates improving the Transparency of training data information, requiring Regulatory Authorities to urge all subjects in the AI lifecycle to proactively disclose data and training information.

Our country should draw on the European Union's practical experience to strengthen the construction of the training data Transparency system.At the legislative level, the data processing Transparency Obligations of Generative Artificial Intelligence service providers are clearly defined, requiring them to fully disclose to users the purposes, scope, methods, and potential risks of data usage. A full-process record system for data processing is established to achieve traceable management of data processing activities. Meanwhile, efforts are increased in the research and application of data anonymization and De-identification Technology to balance the needs of data security protection and value extraction. Enterprises are guided to build internal Data Governance systems, enhancing data quality and Transparency through strengthened self-supervision, and promoting the formation of a self-regulatory Data Governance ecosystem within the industry.

4.3 Analysis and Reference of United States Data Utilization and Supervisory Model

United States adopts a pragmatic attitude prioritizing industry development in Generative Artificial Intelligence Data Governance, relying mainly on Industry Self-regulation and Self-regulation. In terms of public personal data processing utilization, United States federal and state legislation excludes publicly available information from the definition of Personal Information, holding a positive attitude towards the circulation and utilization of publicly available personal information. At the federal level, United States has established a Unified Open Platform (data.gov), which possesses advantages such as standardized specifications, large

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

volume, rich variety, and frequent updates, providing a good resource foundation for AI Data Training. However, this relatively lenient supervisory model in United States also brings some issues, such as the lack of comprehensive data privacy laws leading to some data processing activities possibly being in a regulatory void, and significant differences in data processing Standards across different industries and enterprises, which may easily lead to data misuse and unfair competition.Our Country can draw on the experience of United States in the construction of public Data Openness platforms, further improving our Country's Data Openness system, enhancing the quality and usability of public data, and providing richer data resources for the development of Generative Artificial Intelligence. However, at the same time, excessive Reliance on Industry Self-regulation leading to insufficient supervisory should be avoided. Our Country should strengthen the construction of Law and regulation, clarify the basic principles and bottom-line requirements of data processing activities, and strictly regulate data processing behaviors involving Individuals' privacy, Country security, and social public interest. While encouraging industry innovation, a sound supervisory mechanism should be established, strengthening the supervision and inspection of the data market, ensuring the legal, fair, and orderly circulation and utilization of data.

5 Recommendations for Improving Copyright Protection in the Context of Generative AI Data Openness

5.1 Clarifying the Principles of Copyright Attribution

Establish a model for allocating copyright based on the extent of each party's contribution to the generative AI creation process. Data providers supply the original materials, algorithm developers establish the rules for generating content, model trainers invest computational resources and technical effort, and users may filter or adjust the final generated content. By quantifying the contributions of each party, their respective shares in copyright can be determined, thereby addressing the ambiguity of copyright attribution to some extent and incentivizing active participation in the development of generative artificial intelligence.

Consider establishing a special rights system for generative AI systems, similar to the special rights protection for databases. This right would be independent of traditional copyright and grant certain rights protection to entities that systematically integrate data and generate new content. Such special rights could acknowledge the uniqueness and creativity of AI-generated content without affecting the original copyright of data providers, while also facilitating the management and transaction of AI-related intellectual property rights in an open data environment.

5.2 Expand the scope of fair use

Clearly define reasonable scenarios for using copyrighted data in the field of generative AI for purposes such as training models, improving algorithm performance, and promoting technological innovation. Data use within a certain volume for non-commercial research and development purposes may be considered fair use. Require users to label and explain the source of the data to protect the rights of data providers and ensure the traceability of works. Incorporate public interest into the consideration of fair use. When the application of generative AI can bring significant public interest, appropriate relaxation of restrictions on data use is permitted.

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

5.3 Simplify data usage authorization mechanisms

Establish a centralized, standardized authorization platform to integrate authorization information for various data resources. Data providers can specify the scope, conditions, and authorization fees for data usage on the platform, while data users can conveniently obtain the necessary data authorization through the platform, thereby reducing authorization transaction costs. The platform can provide standardized authorization agreement templates to ensure transparency and legality in the authorization process, while also using technical means to achieve automated management of authorizations, thereby improving authorization efficiency.

Explore new authorization models. Provide tiered authorization schemes based on the different purposes of data users, such as research, commercial applications, and creative development. For data use for research purposes, more lenient authorization conditions can be provided to promote knowledge innovation; for commercial applications, reasonable authorization fees and restrictions can be determined based on their profit scale and impact. Additionally, consider incorporating some concepts from open-source authorization models to encourage data providers to open data under certain conditions, promoting data sharing and innovation.

5.4 Standardize rules for the use of publicly available data

The government should strengthen the management of public data, establish detailed rules and procedures for data openness, and promote data standardization. By establishing a unified data openness platform, the government should increase the openness of public data. Incentivize enterprises and all sectors of society to participate in the development and application of public data to promote data sharing and circulation, and enhance the efficiency of data utilization. The government should also establish a public data openness directory, detailing data classification, format, update frequency, etc., to facilitate easier access and utilization of data by enterprises and the public. Establish a data usage authorization mechanism to ensure that data users sign usage agreements when accessing data, committing to comply with relevant rules.

Unify standards for data anonymization, strictly adhere to the principle of "reasonable anonymization," and clarify technical requirements and evaluation criteria. For data anonymization, specify the use of irreversible encryption algorithms and other technical measures to ensure that anonymized data cannot be reversed back to its original form. At the same time, ensure that the anonymized data retains sufficient usability for subsequent data analysis and processing.

When enterprises use data legally for research and development or applications, if they encounter claims of rights, they should pay fees or provide compensation in accordance with regulations. This protects the rights of data providers while also promoting innovation and progress in artificial intelligence technology. During implementation, it is also necessary to clarify specific details such as the scope of application of the safe harbor and compensation standards to prevent disputes and abuse. For companies using publicly available data for training, if no claims are received within the specified timeframe, such use shall be deemed lawful; once a claim is received, appropriate compensation standards should be determined based on factors such as data usage and market value.

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

5.5 Establish a policy framework for the open access of generative AI data

Generative artificial intelligence is the core driving force of the new round of technological revolution and industrial transformation, and it is a strategic layout concerning the enhancement of national competitiveness. In the development process of China's information industry, industrial policies have played an important role. However, generative artificial intelligence is developing rapidly, and the current policy system urgently needs to be upgraded and improved. Although China has already introduced relevant policies, there are still issues such as insufficient policy system and targeting, and difficulties in implementing some policies in promoting the development of the generative artificial intelligence industry.

To achieve the healthy development of the generative AI industry, the legalization of industrial promotion policies is imperative. From a national strategic perspective, it is necessary to clarify the industry's development status and objectives and incorporate them into the key content of the national science and technology development plan. When formulating the "Artificial Intelligence Law" in the future, it is essential to systematically integrate and clarify relevant promotional provisions to establish a solid legal foundation for industry development. Local governments and industry departments should actively explore pilot initiatives based on actual conditions, formulate detailed policy measures, and establish a multi-level, comprehensive policy support system to stimulate corporate innovation and drive rapid industry development.

References

[1] Zhu, Renzhen. (2024). An Exploration of Copyright Issues in Generative AI Content Creation in the Context of ChatGPT (Master's thesis, Jingdezhen Ceramic University). Master'shttps://link.cnki.net/doi/10.27191/d.cnki.gjdtc.2024.000498doi:10.27191/d.cnki.gjdtc.2024.000498.

[2] Xu Jie. (2024). Analysis of the Attribution of Copyright for Content Generated by Artificial Intelligence. Legal System Review, (07), 67-69.

[3] Liu, Xiaochun. (2024). "Non-copyrighted use" in generative artificial intelligence data training and its legal justification. Law Forum, 39(03), 67-78.

[4] Tan Wei. (2024). Exploring Pathways to Mitigate Copyright Infringement Risks Associated with Generative AI. Communication and Copyright, (14), 107-110. doi:10.16852/j.cnki.45-1390/g2.2024.14.030.

[5] Ding Meng & Liu Yingze. (2024). A Study on Copyright of Generative Artificial Intelligence with a Focus on Data Protection. Journal of Hebei Open University, 29(05), 78-82. doi:10.13559/j.cnki.hbgd.2024.05.017.

[6] Wang, C. (2024). Copyright risks and mitigation strategies in the development of generative artificial intelligence: A case study of ChatGPT. Journal of Huaiyin Normal University (Philosophy and Social Sciences Edition), 46(02), 139-145+154+215-216.

[7] Cao Xinming & Fan Ye. (2024). Research on reasonable usage rules for generative artificial intelligence data training. China Copyright, (04), 20-35.

[8] Ding Daoqin. (2024). Data legal issues in the training phase of generative artificial intelligence and legislative recommendations. Administrative Law Research, (06), 16-28.

[9] Zhang Huibin & Li Junhao. (2024). International Comparison of AIGC Copyright Protection in the Era

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

of Intelligent Creation and China's Practice. Journal of International Economic Law, (03), 64-79.
[10] Zhang Ping. (2024). Institutional Responses to Copyright Issues in Generative Artificial Intelligence. Southern Media Research, (05), 21-31.
[11] Li Mengfei. (2024). The Attribution of Rights and Protection Pathways for Generative AI-Created Works in the Field of Copyright. In: Legal Practice Journal, Vol. 2, 2024 — Collection of Essays on Legal Guarantees for New Industrialization (pp. 192-202). East China University of Political Science and Law; doi:10.26914/c.cnkihy.2024.024079.
[12] Zhang Xinbao & Bian Long. (2024). Copyright Protection of Training Data for Generative Artificial Intelligence. Jingchu Law Review, (05), 77-90.
[13] Liu, Yanhong. (2023). The Three Major Security Risks of Generative Artificial Intelligence and Legal Regulation: Taking ChatGPT as an Example. Dongfang Law Review, (04), 29-43. doi:10.19404/j.cnki.dffx.20230606.003.
[14] Sun, Xiaoqi. (2023). Research on Copyright Issues Related to AI-Generated Content. China Copyright, (06), 51-59.
[15] Sun Zhengliang. (2019). An Analysis of Copyright Issues Related to Content Generated by Artificial Intelligence. Tsinghua Law Review, 13(06), 190-204.

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)