ISSN: 2710-0170

Research on Copyright Infringement Issues in Generative Artificial Intelligence

Shuyi Wu [1], Ji Zhen [1], Jingyi Tong [2], Jie Song [3]

[1] School of Law, Anhui University of Finance & Economics, Bengbu Anhui 233030, China;

[2] Institute of Finance and Public Management, Anhui University of Finance & Economics, Bengbu Anhui 233030, China;

[3] School of International Trade and Economics, Anhui University of Finance & Economics, Bengbu Anhui 233030, China.

Abstract

Generative Artificial Intelligence (GAI) technology has achieved rapid development since the release of ChatGPT and Sora, with significant breakthroughs in domestically developed models. However, the copyright legal issues it raises are becoming increasingly prominent. This paper, based on the principles of GAI generation and the originality standard of works, argues that GAI training data meeting the " originality " standard possesses the attributes of copyrightable material. China's current Copyright Law and related legal norms are insufficient to address practical problems such as disputes over the copyright status of GAI training data, the inapplicability of moral rights, ambiguity of rights holders, and difficulties in pursuing infringement liability. Drawing on international legislative experience, this paper proposes that China should clarify the protectability of GAI training data by adding copyright exception clauses, attributing its copyright to users while appropriately restricting moral rights. Simultaneously, it should distinguish the duty of care and infringement liability of platform developers and users, constructing a copyright regulatory system that aligns with the development of GAI technology, achieving a balance between rights protection and innovation in the literary and artistic fields, and promoting the sustainable development of generative artificial intelligence.

Keywords

Generative artificial intelligence, Copyright protection, Infringement liability.

1. Introduction

In November 2022, OpenAI launched ChatGPT, a generative artificial intelligence (GAI) model. GAI officially empowers the creation of literary and artistic content. Subsequently, the company launched Sora, which can convert text into video. GAI has achieved a leap from "being able to speak" to "being able to perform" visualization. Chinese technology companies have localized and transferred GAI, and software such as Kimi and Doubao have emerged. It is worth mentioning that Hangzhou Deepseek Artificial Intelligence Basic Technology Research Co., Ltd. developed Deepseek , which has the performance of top-tier pre-trained large models in the world with only extremely low computing power cost . GAI has achieved vigorous development from "being able to speak", "being able to perform" to "being able to think". However, as a disruptive technology, generative artificial intelligence involves complex technical algorithm logic. In addition, due to reasons such as regulatory lag [0], technical ethics issues have emerged. At present, the Civil Code of the People's Republic of China , the Copyright Law of the People's Republic of China and related judicial interpretations are difficult to regulate the copyright issues brought about by GAI training data . Specifically, the protectability of GAI as a copyright

ISSN: 2710-0170

object is controversial, and it is difficult to apply the moral rights system . The problems of unclear copyright subject and difficulty in holding infringers accountable are prominent. In light of this, this paper analyzes the copyright infringement issues involved in the GAI training process, affirms that GAI training data that meets certain conditions constitutes copyright protection, and draws on international experience to design practical and feasible strategies for the long-term development of GAI.

2. Protectability Analysis of Generative Artificial Intelligence Training Data

According to the generally accepted theory in academia, a work refers to an intellectual achievement in the fields of literature, art and science that has originality and can be expressed in a certain form. According to the analysis of the generation principle of GAI, the user input instructions reflect the thinking of the human brain. Even if the same instructions are input, different training data may be generated. It has randomness and originality, which meets the definition of "work". That is, it has the characteristic of "originality". GAI training data can be used as the object of copyright protection. In judicial practice, the first judgment made by the Beijing Intermediate People's Court to protect artificial intelligence generated images [2]shows that affirming that GAI training data can constitute the object of copyright is an inevitable trend.

2.1. Principles of GAI Training Data Generation

GAI (Generative Artificial Intelligence) is an artificial intelligence technology that automatically generates content based on prompts in a natural language conversational interface [3]. The operation of GAI mainly goes through three stages, including the preparation stage, the computation stage, and the generation stage [4]. In the preparation stage, researchers of generative artificial intelligence mainly obtain training data in five ways: collecting relevant information on their own; using publicly available datasets; extracting data from the Internet using technical means; purchasing data resources from third-party institutions; and generating the required data through simulation [7]. After the user issues an instruction, GAI uses these five methods, but is not limited to these five methods, to collect data and prepare the required data for the computation stage. In the computation stage, GAI uses complex algorithms to integrate and reorganize the prepared data, including the cumulative fusion application of technologies such as GAN , CLIP , Transformer , Diffusion , pre-trained models, multimodal technology, and generative algorithms . In the final generation stage, GAI provides the required training data according to the user's command.

2.2. GAI training data that meets specific requirements should constitute the subject matter of copyright protection.

Works eligible for copyright protection must meet the requirement of "originality." According to the Supreme People's Court and internationally recognized viewpoints , "originality" refers to a work meeting the standard of distinguishing itself from other copyright holders' styles and expressions, while "creativity" only requires a minimum level of originality. In the era of weak artificial intelligence, GAI generates training data based on user instructions. Due to the high degree of randomness in its algorithms, even if users input the same instructions, different content is highly likely to be generated. GAI can generate training data that is distinct from other works and possesses different characteristics, thus meeting the "originality" requirement for copyright protection. Furthermore, the fundamental purpose of copyright protection is to incentivize innovation in literature, art, and science. GAI is essentially just a tool to assist human creation. Using this tool, recognizing its status, and protecting its legitimate rights can greatly incentivize innovation. Therefore, when GAI training data does not infringe on the rights of

ISSN: 2710-0170

other copyright holders and possesses the core "originality" standard of copyright, it should be recognized as copyrightable material, and its legitimate rights should be protected by copyright.

3. Challenges in the Development of Generative Artificial Intelligence

GAI has achieved vigorous development from "being able to speak," "being able to perform," and then "being able to think," playing an important role in the organization and application of information and becoming a new engine for the development of the digital economy and society. However, its vigorous development has also brought new problems. The traditional objects of copyright protection are inherently prone to "easy infringement, difficult rights protection . " Currently, the protectability of GAI as an object of copyright is controversial, making it difficult to apply the moral rights system . New problems such as the ambiguity of copyright holders and the difficulty in pursuing accountability for infringement make it difficult to protect the legitimate interests of copyright holders, thus hindering innovation in the fields of literature and art. It is necessary to make a positive response from the legal perspective.

3.1. Missing natural person attribute requirement

Although the generation of GAI training data requires instructions from natural persons, it mainly relies on the complex GAI algorithm and lacks the natural person attribute requirement. On the one hand, the prevailing theory does not recognize its status as a copyright object. The classic American "monkey selfie" case [8]denied the status of works created by non-natural persons. Photographer David Slater's camera was snatched by a monkey and more than a hundred photos were taken. When Slater profited from these photos, he was sued by an animal protection organization, which believed that Slater had infringed on the monkey's copyright.[9] The court ultimately held that animals cannot enjoy copyright and emphasized that only human creations can be protected by copyright law. The prevailing theory represented by the case requires that the work must be created by a natural person, while GAI training data is actually generated by GAI and is essentially a "work" given by the algorithm.[10] Natural persons only play a limited role in inputting passwords. On the other hand, GAI is difficult to apply the moral rights system. The author's moral rights are closely linked to the natural person, cannot be separated from the author's person , cannot be transferred or inherited, and are not subject to the [5]. As a product of technology, GAI training data cannot be subject to the rights of the algorithm itself. The creation process is relatively simple, yet the algorithm enjoys too many rights, which has aroused dissatisfaction among traditional creators and violated the principle of fairness.[11]

3.2. Unclear copyright holder

When GAI training data meets certain conditions and is granted copyright protection, disputes over ownership of rights arise. First, according to the contribution theory, the right belongs to whoever makes the greatest contribution; in this case, it should belong to the developer. However, this creates a "double protection" phenomenon: the copyright of the algorithm itself already belongs to the developer. If the training data, as a technological product, also belongs to the developer, the developer will benefit twice. This overprotection can inhibit users' creative enthusiasm.[12] Second, according to the tool theory, GAI is merely a tool to assist users in creation; in this case, the right should belong to the user.[13] However, the user's role is relatively small, merely providing a few short commands, which unduly lowers the already low threshold for copyright protection, resulting in inadequate protection.[14] Finally, according to the fictive theory, a legal personality is created for GAI , and the copyright of its training data belongs to the public interest. This can dampen companies' enthusiasm for technological development and significantly reduce users' creative enthusiasm.[15]

ISSN: 2710-0170

3.3. The infringing party is difficult to identify

First, the GAI "black box" problem is significant. The complex algorithm and operating logic bring about the "black box" problem - the generation process of its training data lacks transparency and interpretability , builds a high barrier to public understanding, and even some experts find it difficult to understand and evaluate. The technical level is still difficult to achieve a certain degree of transparency, which makes it difficult for legal supervision to go deep . Second, infringement may occur at any stage of GAI creation. In the preparation stage , copyright-protected works may be improperly captured. In the massive digital resources, demanding the legal authorization of each copyright owner, or being unable to obtain high-quality training digital resources due to high data transaction costs, will restrict the innovative development of artificial intelligence. On the contrary, it will infringe on the legitimate rights and interests of copyright owners. In the calculation stage , infringing works will be mixed into the training data to be generated. In the generation stage , training data that infringes on the rights and interests of other copyright owners may be generated due to the algorithm or user password. Finally, the generation of training data is random. Although developers and users have some control over the design and input data , the specific form of expression is not directly determined by human behavior [0]. Complex algorithms play a dominant role. Before determining the responsible party, the complexity of the technology exacerbates the difficulty of assigning blame, and the potential for infringement at every stage makes localized supervision inadequate. When formally identifying the responsible party, each party has its own shortcomings. If the algorithm is held accountable, the algorithm itself cannot bear responsibility; if the source is traced back to the platform, excessively strict liability unduly increases its duty of care, thereby inhibiting technological innovation and hindering the development of high-tech industries; if the user is held accountable, on the one hand, users who only input passwords have limited influence, and on the other hand, users without the intention to infringe are "forced" to infringe due to the black-box algorithm, which does not conform to the principle of consistency between subjective and objective factors in accountability.

4. The Construction Path of Generative Artificial Intelligence Legal System

4.1. Exceptional legislation clarifies the object of rights

British jurist Maine once said, "Once a law is enacted, it is already outdated ." With the rapid development of GAI and the disruptive changes it has brought to society, the traditional objects of copyright protection are no longer suitable for the current reality. Legislation should be changed according to the changes in reality. Foreign countries have already responded to this reality. The "Text and Data Mining (TDM) Exception Clause" [6]of the EU's Digital Single Market Copyright Directive has clearly granted research institutions the right to conduct data mining under specific conditions. Japan has introduced a more flexible "soft exception clause" in its copyright law. Combining foreign judicial experience, China's Copyright Law can add similar clauses when it is amended. Based on the principle of traditional objects of protection, an exception is set for GAI training data that meets specific "originality" characteristics, in addition to the principle. This would make up for the lack of natural person attributes of GAI and make the legislation clear that GAI training data that meets specific requirements can constitute the objects of copyright protection.

4.2. Ownership of rights to the user

Once the rights are clearly defined, the question of ownership arises. On the one hand, it is inappropriate to assign the rights to the developers. Developers have already recouped their investment; granting copyright again would be detrimental to fair competition and create a "Matthew effect." On the other hand, it is inappropriate to assign the rights to the public interest.

ISSN: 2710-0170

Protecting copyright is for social development, but premature intervention could stifle users' creative passion. In conclusion, the copyright of GAI training data should belong to the users. Although users did not substantially design the AI's creative algorithm and their contribution to creation is extremely limited, they have already paid a reasonable price and are the primary creators. Copyright protection would incentivize their creation and promote innovation in literature, art, and other fields. However, given the limited role users play in generating GAI training data, legislation should, while equally protecting the economic rights of GAI works, appropriately limit the moral rights of GAI authors to balance the interests of traditional creators and GAI training data users.

4.3. Liability for Tort

4.3.1. Platform Developer Responsibilities

"Technological neutrality" cannot be used as a reason for platforms to be exempt from liability . Once infringement occurs, the responsibility should be traced back to the developers . The GAI industry has high professional barriers, and platform developers should bear a greater duty of care. However, this does not mean they bear unlimited strict liability; rather, their liability for infringement is determined based on the fulfillment of their duty of care. According to the "notice-and-takedown" rule (also known as the safe harbor rule) stipulated in Article 1195 of the Civil Code of the People's Republic of China , and considering the actual costs of enterprise development, platform developers should actively assume a duty of care based on the standards of general service providers in the same industry. When copyright disputes are brought to court, if the developer can prove that they have taken necessary preventative measures beforehand, made necessary technical disclosures and explanations to regulatory authorities, and promptly deleted or stopped the dissemination of infringing content afterward, they will not be liable. If they allow infringement to occur, they should bear liability for infringement.

4.3.2. User Responsibility

As the direct creators of GAI works, users can input different commands according to their subjective needs. However, when they input commands with an infringing intent, they infringe on the copyrights of others. On the one hand, if a user inputs commands normally, but infringement occurs due to the platform's failure to act as a "gatekeeper," the user should not bear liability for infringement since they had no subjective intent to infringe. On the other hand, if a user inputs a targeted command that induces the algorithm to output works that infringe on the copyrights of others, the user should bear liability for infringement. Furthermore, because GAI works are difficult to distinguish from works in the ordinary sense, users should clearly indicate that their works are GAI works.

5. Conclusion

General Secretary Xi Jinping pointed out: " Artificial intelligence is an important driving force for a new round of technological revolution and industrial transformation, and will have a profound impact on global economic and social development and the progress of human civilization ." In the stage of weak AI, GAI is merely a tool to assist human creation , and its development should always be encouraged with a prudent and inclusive attitude . At the same time, GAI, as a new type of productive force, has enormous development potential and should rightfully constitute a subject of copyright protection. However, problems such as the lack of natural person attributes, unclear copyright holders, and difficulty in identifying infringers hinder its development. To rebuild trust in technology and incentivize innovation in literature and art, we should learn from international experience. This could involve clarifying its subject status through exceptional legislation, attributing copyright to users, and pursuing

ISSN: 2710-0170

infringement liability in a tiered, situational, and phased manner to address practical problems, achieve a balance of interests among multiple parties, and promote the development of GAI towards enhancing the common well-being of all humanity.

Acknowledgements

This work is supported by Anhui University of Finance & Economics 2025 Undergraduate Research innovation fund project fund, Project number: XSKY25221.

References

[1] Yang Jianwu, Luo Feiyan. Hierarchical and Classified Management: Content Risks and Legal Regulation of Generative Artificial Intelligence [J]. Journal of Kunming University of Science and Technology (Social Sciences Edition), 2025, 25(06): 13-22.

[2] See the Beijing Internet Court Civil Judgment , (2023) Jing 0491 Min Chu No. 11279_Global

[3] See UNESCO's Guide to Generative Artificial Intelligence in Education and Research.

[4] Ma Yunan . Risks and Governance of Generative Artificial Intelligence—Taking ChatGPT as an Example [N]. Chinese Social Sciences Daily, 2024-05-15 (7).

[5] Yuan Cheng. A Study on the System of Reasonable Use of Authors' Moral Rights [D]. Central China Normal University, 2017.

[6] Articles 3 and 4 of the EU's Digital Single Market Copyright Directive specifically establish exceptions for text and data mining, explicitly stipulating that text and data mining for scientific research purposes can be conducted without the authorization of the rights holder.

[7] Gao Zejin . Pandora's Box: Sources, Uses and Governance of Artificial Intelligence Training Data — A Rooted Study for 100 AI Developers [J]. Journalist, 2022(1): 86-96.

[8] (See Naruto v. Slater (888 F. 3d 418).)

[9] Zhao Lixin, Liu Jiaying . Determination of tort liability for generative artificial intelligence behavior [J]. Journal of Hebei Normal University (Philosophy and Social Sciences Edition), 2025, 48(05): 147-156.

[10] Liu Cheng. Copyright Protection of Generative Artificial Intelligence Data Training and Its International Experience [J]. Chongqing Social Sciences, 2025, (07): 6-17.

[11] Wang Ruoyu , Hu Shensong . Data traceability dilemma and collaborative governance path for copyright determination of content based on generative artificial intelligence [J]. Journal of Kunming University of Science and Technology (Social Sciences Edition), 2025, 25(06): 32-40.

[12] Wang Liming. Attribution Principles and Fault Determination in Generative Artificial Intelligence Torts [J]. China Law Review, 2025, (04): 15-30.

[13] Song Yunbo, Lu Yang . Dilemmas, Models and Legislative Improvement of Copyright Determination for Generative Artificial Intelligence Works [J]. Digital Economy and Rule of Law, 2024, (02): 49-67+243.

[14] Feng Xiaoqing, Guo Chang. Copyright Infringement Liability and Hierarchical Governance of Generative Artificial Intelligence Platforms — Reflections Based on the Hangzhou Ultraman Case [J]. Digital Rule of Law, 2025, (02): 56-75.

[15] Nie Hongtao , Chen Yifan. Copyright of Generative Artificial Intelligence Works: Ownership and Institutional Construction [J]. Hainan Finance, 2024, (03): 77-87.