Asian Journal of Law and Policy
Vol 5 No 3 (December 2025) eISSN: 2785-8979

Legal Challenges and Evidentiary Barriers in Determining Copyright Infringement by Generative AI: A Taiwan-Centred Analysis

Jia-Yi Wu
National Cheng Kung University, Taiwan
kellywu871030@gmail.com
ORCID iD: 0009-0005-0006-8069
(Corresponding author)

ABSTRACT

This study examines the evidentiary challenges in copyright infringement litigation involving generative AI technologies under Taiwan's legal framework. Drawing from the operational mechanisms of large language and diffusion models, it explores the difficulty copyright holders face in proving substantial similarity or unauthorised reproduction when AI developers refuse to disclose training datasets. The paper analyzes two high-profile US cases: *Andersen v Stability AI*, where the court partially dismissed the plaintiffs' claims due to insufficient factual allegations regarding compressed copies and third-party use, and *The New York Times v OpenAI and Microsoft*, in which plaintiffs submitted outputs from ChatGPT and Browse with Bing that closely resembled original copyrighted articles, suggesting potential infringement of reproduction and derivative rights. These cases illustrate both the legal uncertainty and the potential for novel evidentiary strategies. The paper argues that prompt engineering—crafting input commands to provoke infringing outputs—may assist plaintiffs in building stronger prima facie cases. Finally, the paper proposes legislative reform by introducing a statutory licensing scheme specifically tailored to AI-related uses in Taiwan, aiming to reduce the evidentiary burden on authors and ensure fair compensation.

Keywords: Generative AI; Copyright infringement; Evidentiary challenges; Taiwan

Received: 15 June 2025, **Accepted:** 22 August 2025, **Published:** 1 December 2025

MULTIMEDIA UNIVERSITY
(2025) 5(3) Asian Journal of Law and Policy 369–383
https://doi.org/10.33093/ajlp.2025.19
© Universiti Telekom Sdn Bhd. This work is licensed under the Creative Commons BY-NC-ND 4.0 International License.
Published by MMU Press. URL: https://journals.mmupress.com/ajlp

1. Introduction

1.1 Background and Research Motivation

The copyright system was established in the wake of the widespread adoption of printing technology, which enabled the rapid and inexpensive reproduction of literary and artistic works in tangible form. These physical embodiments of creativity allowed for the swift distribution of content through ownership transfers. However, creators did not proportionately benefit from the increased efficiency in distribution. To address this imbalance, and to incentivize creative endeavours while ensuring fair allocation of social resources, copyright was introduced as a form of intangible property protection—providing economic rights to creators that could not be secured through traditional property law frameworks.[1]

Crucially, the copyright regime is not solely aimed at safeguarding private interests. As stated in Article 1 of the Taiwan Copyright Act, its legislative purpose is 'to protect the rights and interests of authors with respect to their works, balancing different interests for the common good of society, and promoting the development of national culture'. Thus, the Copyright Act seeks to strike a balance between private rights and the public good in the pursuit of social justice.[2]

With the launch of ChatGPT in late 2022, access to information has become more rapid and convenient. Unlike traditional search engines that merely redirect users to existing websites, generative AI tools like ChatGPT search internal databases using probabilistic and logical models to produce content—text, audio, images, or video—autonomously. These machines often outperform human capabilities in generating expressive works, raising concerns over AI's potential to replace human creators. These concerns are not unfounded, as Generative AI has reached a level where its outputs fully encompass the expressive forms traditionally achievable only through human creativity. For instance, AI-generated vocal imitations—such as 'AI Stefanie Sun'—have attracted massive attention on Chinese web platforms, with covers of songs like Red Bean, A Game, A Dream, and Silence rendered in a voice nearly identical to the original singer's. In addition, AI technologies have been used to replace actors and scriptwriters, contributing to the tensions underlying the recent Hollywood strikes. AI-generated visual art has also raised concerns, as demonstrated by the controversy over Chaos in the Dragon Palace, a prize-winning piece created by a Taiwanese vocational high school student using generative drawing tools. These examples illustrate how the emergence and rapid development of generative AI not only encroach upon the creative space traditionally occupied by humans but also pose significant risks of copyright infringement.

---

[1] Chung-Hsin Chang, 'The Origin and Development of Compulsory Licensing—From the Perspective of Copyright Law' (The 6th International Symposium on Intellectual Property Rights between the European Union and East Asia, June 2015).

[2] Chang (n 1).

Nevertheless, whether human creators can claim their rights under the existing copyright regime introduces a critical tension between 'rights protection' and 'cultural development'. Although generative AI utilises original works from databases, its outputs are the result of integrating and modifying such materials. As a result, they often do not constitute 'substantial similarity' to the originals and therefore may not constitute 'plagiarism'. Furthermore, the act of collecting and reproducing original works in training datasets may or may not constitute infringement of reproduction rights under Article 3(1)(5) of Taiwan's Copyright Act—this remains legally unsettled.

In this case, this study aims to explore whether the operation of generative AI under the development of modern technology constitutes copyright infringement and, if so, how authors can fulfil the evidentiary burden in litigation. This paper begins by outlining the mechanisms behind generative AI, forming the basis for analysing infringement and fair use. It then discusses the evidentiary challenges in litigation and proposes potential solutions, followed by legislative recommendations.

1.2 Methodology

This study adopts a hybrid doctrinal and comparative-empirical methodology. The doctrinal component involves a systematic analysis of Taiwan's Copyright Act, relevant judicial decisions, and administrative interpretations to determine how the existing legal framework addresses the evidentiary burden in cases involving AI-generated outputs. Special attention is given to Articles 3, 22, and 28 of the Act, which govern the rights of reproduction and adaptation, and to Articles 44–65 concerning statutory exceptions and fair use.

In parallel, the comparative analysis draws upon recent developments in US case law—specifically *Andersen v Stability AI Ltd* and *The New York Times Company v Microsoft Corporation and OpenAI LP*—to illustrate how evidentiary challenges are currently addressed under the common law approach, and to assess the potential relevance of these approaches to Taiwan's civil law system. These cases were selected based on their early judicial treatment of generative AI copyright claims and their influence on global copyright discourse.

The study also incorporates functional legal analysis, examining the technical architecture of generative AI models (eg, LLMs, diffusion models), including training mechanisms, prompt engineering techniques, and model opacity, to evaluate their impact on traditional infringement doctrines such as 'access' and 'substantial similarity'. To bridge law and technology, the paper critically assesses whether targeted prompts—when used to elicit infringing outputs—can serve as prima facie evidence of unauthorised copying.

Finally, this study applies policy-oriented legal reasoning to develop legislative recommendations tailored to Taiwan's civil law system, including the introduction of a statutory licensing regime to rebalance evidentiary burdens and secure fair compensation for authors whose works are used in large-scale AI training without prior authorisation.

2. Operation of Generative AI

2.1 General Overview

Generative AI differs from traditional artificial intelligence in that it does not merely output existing data. Rather, it 'learns' from vast datasets and, through pre-programmed computational systems, produces new content that differs from the original data. This process involves three core technologies: machine learning, neural networks, and large language models (LLMs).[3] Each is briefly introduced below.

2.1.1 Neural Networks

IA neural network is a type of computational model inspired by the biological nervous system, in which artificial neurons—the smallest processing units—are interconnected to transmit information. Through these connections, the network is capable of processing complex data across multiple layers. This layered structure enables the network to perform nonlinear computations, automatically detect patterns in data, and make predictions or generate outputs based on those patterns. Without relying on explicitly defined instructions for each task, neural networks can learn to handle diverse and complex problems by generalising from the data they are trained on. In generative AI systems, such neural networks serve as the fundamental architecture behind deep learning and are central to how these systems produce coherent and contextually relevant outputs.

2.1.2 Large Language Models (LLMs)

Large language models are systems trained to process human language, capable of learning linguistic rules through large-scale corpora and generating coherent natural language outputs. These models are widely applied in natural language processing tasks such as question-answering systems, summarisation, machine translation, and content creation. Its underlying dataset has exceeded one billion data points, enabling it to generate text and be combined with other generative models.

2.2 Practical Applications of Generative AI

2.2.1 Stable AI

Stable AI is a tool that generates images based on textual prompts, primarily using a diffusion model for machine learning. Since AI models process only binary data (0s and 1s),

[3] 'What Is Generative AI? Understanding the Principles and Applications of Generative AI' (*Preface AI Blog*) <https://www.preface.ai/blog/what-is-generative-ai/>.

each image input is first encoded through an autoencoder to assign pixel values and compute probabilistic distributions.[4]

Next, cosine similarity is calculated—values closer to 1 indicate a higher match between the image and the prompt. The system then applies a forward diffusion process (adding noise) and a reverse diffusion process (denoising), progressively transforming the image from clarity to blur and back to clarity. This process often results in output images with higher quality than the originals.

Finally, the generated images—tailored to the user's prompt—are synthesised through the above procedures to create outputs that differ from any image in the training dataset. The model deals with complex mathematical problems and algorithms, reflecting the use of deep learning technology.

2.2.2 ChatGPT

ChatGPT,[5] developed by OpenAI, operates as a language generation tool grounded in a Large Language Model (LLM) architecture known as the Generative Pre-trained Transformer (GPT), which is itself a prominent deep learning framework utilizing a neural network structure called the 'Transformer'; this framework, designed to enable advanced natural language processing capabilities, undergoes a two-stage training process whereby, in the initial pre-training phase, the model is exposed to large-scale publicly available text corpora (such as websites, books, and encyclopedic sources) to perform unsupervised learning through next-word prediction, thereby enabling it to internalize semantic logic, syntactic structure, and contextual relationships in natural language, and subsequently, in the fine-tuning phase, the model is subjected to supervised learning based on human-labeled datasets in order to improve task-specific performance—with additional optimization implemented through Reinforcement Learning with Human Feedback (RLHF), which aligns the model's outputs with ethical norms and user expectations; operationally, ChatGPT converts input sentences into vector representations and processes them through multi-layer Transformer encoding and decoding structures, estimating the most probable subsequent word via statistical modeling, and consequently generating coherent and contextually appropriate language output, such that, as a typical application of deep learning characterized by high parameter volume and computational complexity, it relies entirely on parameterized mappings derived from machine learning rather than any rule-based or direct data retrieval from static databases.

---

[4] Chen Jiajun and Xu Zhenggan, 'Midjourney and Stability AI Diffusion Model Automatic Generation Raises Copyright Infringement Suspicion?–Graphics and Imagery' (*STPI IKnow Center*, 21 June 2023) <https://iknow.stpi.narl.org.tw/post/Read.aspx?PostID=19796>.

[5] Chen Jiajun and Xu Zhenggan, 'Media Giants Sue ChatGPT for Copyright Infringement–New York Times v Microsoft and OpenAI' (*STPI IKnow Center*, 2 April 2024) <https://iknow.stpi.niar.org.tw/post/Read.aspx?PostID=20588>; 'What Is ChatGPT? Definition and Origins of ChatGPT' (*CAP Networking Camp*, 31 August 2024) <https://www.cap.com.tw/modules/news/article.php?storyid=52>.

3. Potential Copyright Infringement and Fair Use in Generative AI

3.1 The Possibility of Copyright Infringement

3.1.1 The Legal Standard for Infringement

According to the jurisprudence of Taiwan's Supreme Court, the concept of 'plagiarism' once encompassed both 'idea plagiarism' and 'expression plagiarism'.[6] However, recent decisions have adopted a more stable and narrowed interpretation, confining infringement to cases of expression plagiarism. The Court has stated: 'When determining whether copyright infringement has occurred, the court should examine all relevant circumstances and carefully assess the two essential elements of infringement: namely, access and substantial similarity. Substantial similarity refers not only to quantitative similarity but also to qualitative similarity'.[7]

In other words, the determination of whether an infringement of the right of reproduction or the right of adaptation has occurred must be based on whether the defendant had a reasonable opportunity to access the original work and whether the two works exhibit substantial similarity.

The criteria for evaluating substantial similarity vary by type of work. In the case of artistic works, courts consider the overall impression, atmosphere, and aesthetic appearance of the works. A piecemeal comparison is discouraged in favour of an analysis of the total concept and feel.[8] For literary works, a more analytical, segment-by-segment comparison is conducted, evaluating the quantity and quality of the material used based on objective social standards.[9]

3.1.2 Possible Infringement of Reproduce Right

As discussed in the previous chapter, the outputs of generative AI systems, due to the nature of machine learning, are usually not substantially similar to specific works in their training data. Thus, such outputs are generally not considered direct copies or infringements under the right of reproduction.

However, the training data used to develop generative AI is often collected from various sources, sometimes without authorisation, to reduce costs and enhance commercial potential. Whether this constitutes a violation of the right to reproduce under Article 3(1)(5) of Taiwan's Copyright Act remains debatable.

---

6 Supreme Court Civil Judgment 81-Tai-Shang-3063 (Taiwan, 1992).

7 Intellectual Property Court Civil Judgment 104-Min-Zhu-Shang-Yi-15 (Taiwan, 2015).

8 Supreme Court Civil Judgment 103-Tai-Shang-1544 (Taiwan, 2014).

9 Supreme Court Civil Judgment 97-Tai-Shang-3121 (Taiwan, 2008).

According to the Taiwan Intellectual Property Office (TIPO), if the original training material is protected by copyright, reproducing it for training purposes without the right holder's permission—unless covered by fair use under Articles 44 to 65[10]—could violate the right to reproduce under Article 22.[11]

On the other hand, some scholars argue that simply scanning works optically and storing them in transformed formats (e.g., encoded data) without saving copies in their original expressive form should not constitute reproduction.[12]

This study contends that unless copyrighted material is stored in a manner akin to natural human browsing (i.e., without permanent storage), such use should fall outside the scope of reproduction under Article 10-1,[13] which only protects expressive forms—not ideas or concepts.

3.1.3 Possible Infringement of Adaptation Right

Even when the output of generative AI does not meet the threshold of substantial similarity, it may still raise concerns under the right of adaptation (Article 28 of Taiwan's Copyright Act). If the AI system extracts characteristics from existing works and produces derivative content with recognisable elements, it may be classified as a derivative work requiring authorisation from the original rights holder.

3.2 The Possibility of Fair Use

In Taiwan, the doctrine of fair use is interpreted as an 'exclusion from liability'. That is, even if copyright infringement is established, the court must examine whether the use qualifies as fair under Articles 44 to 65.[14] Since generative AI is typically used for commercial purposes, statutory exceptions (e.g., for education or news reporting) often do not apply. Therefore, the general fairness factors in Article 65(2)[15] become the primary standard.[16]

3.2.1 Purposes and Nature of the Exploitation

Fair use is more likely to be recognized for non-commercial or educational purposes. If the AI output transforms the original work by adding new meaning or functionality, this increases the likelihood of fair use. However, since most generative AI applications are

[10] Copyright Act (Taiwan).

[11] Intellectual Property Office, Ministry of Economic Affairs, Official Letter No 11252800520 (Taiwan, 2023).

[12] Chung-Hsin Chang, 'The Possibility of Fair Use of Generative AI' (*Copyright Note*, 2003) <http://www.copyrightnote.org/ArticleContent.aspx?ID=9&aid=3154>.

[13] Copyright Act (n 10).

[14] Supreme Court Criminal Judgment 94-Tai-Shang-7127 (Taiwan, 2005).

[15] Copyright Act (n 10).

[16] Intellectual Property Court Civil Judgment 107-Min-Zhu-Su-68 (Taiwan, 2018).

commercially driven, they face a higher bar to qualify as fair use. The less the output resembles the original work, the stronger the argument for transformative use.

3.2.2 Nature of the Work

This factor considers the originality, creativity, and availability of the source work. From the perspective of balancing interests, the higher the originality of the utilized work, the narrower the scope for asserting fair use becomes.

Accordingly, courts should assess the originality of the copyrighted work used for model training on a case-by-case basis, in order to determine the likelihood that a fair use defence may succeed.

3.2.3 The Amount and Substantiality of the Portion Exploited in Relation to the Work as a Whole

This factor evaluates both quantitative and qualitative aspects of the use. Copying the core or most expressive parts of a work is less likely to be fair, while using minor or non-essential portions increases the chance of fair use. Since generative AI often extracts style rather than content—and Taiwanese courts generally do not recognise copyright in 'style'—this may favour a fair use argument. However, each case depends on whether the copied part is qualitatively significant.

3.2.4 Effect of the Exploitation on the Work's Current and Potential Market Value

This factor examines whether the use substitutes for the original in the market and harms its commercial value. Most generative AI developers operate with a view toward commercial gain, a motive that frequently results in market substitution or diminishes the economic potential of the original work. Thus, where authors suffer measurable economic harm due to such substitution, the burden arguably shifts to AI developers to demonstrate that their use confers a public interest benefit sufficient to outweigh the resultant loss, thereby warranting protection under doctrines such as fair use or statutory exceptions.

4. Evidentiary Challenges in Determining Copyright Infringement by Generative AI

4.1 The Evidentiary Dilemma in Copyright Infringement

As previously discussed, the output generated by generative AI models often differs from the original copyrighted works stored in their training datasets. As a result, authors typically find it difficult to satisfy the requirement of substantial similarity. In cases where copyrighted works are copied for the purpose of constructing training models used in machine learning, such acts may constitute plagiarism or infringement under copyright law. However, a key challenge lies in the ability of authors to prove such infringement.

Developers of generative AI systems—such as OpenAI, which developed ChatGPT, and Stability AI, the developer of Stable Diffusion—have neither disclosed nor demonstrated any intention to disclose the contents of their training datasets, nor whether such data were lawfully acquired and appropriately licensed. This lack of transparency severely hinders the ability of copyright owners to establish that generative AI has engaged in unauthorised use of protected content, thereby infringing their exclusive rights under copyright law.

4.2 A Procedural Ruling by a US Court—*Andersen et al v Stability AI Ltd et al*

Although no copyright infringement lawsuits involving generative AI have yet been filed in Taiwan, several such cases have already emerged in the United States. A prominent example is the procedural ruling rendered by the US District Court for the Northern District of California in *Andersen et al v Stability AI Ltd et al*,[17] which illustrates the evidentiary difficulties currently faced by authors. In this case, the plaintiffs filed a joint complaint against three generative AI companies—Stability AI, Midjourney, and DeviantArt—alleging that the use of diffusion models by these companies to generate images infringed their copyrighted artistic works.

The plaintiffs alleged that Stability AI released Stable Diffusion as the foundational model for DreamStudio, a tool capable of generating images based on user prompts. The training data for the model was sourced from LAION, to which Stability AI provided financial support for the creation of LAION-Aesthetics, a dataset project that scraped and copied over 5 billion images. These images were stored in the model as compressed copies, an act alleged to constitute direct copyright infringement. Additionally, when users entered the names of specific artists into DreamStudio to generate images in those artists' styles, Stability AI was alleged to have committed contributory infringement.

Although the court found that the plaintiffs had pleaded sufficient facts to support the claim of direct infringement, it noted that contributory infringement requires the plaintiff to demonstrate that the defendant (1) had the right and ability to supervise the infringing activity, and (2) received a direct financial benefit from that activity. Because the plaintiffs failed to allege with particularity how Stability AI stored or distributed 'compressed copies' of the images to third parties, the court concluded that the contributory infringement claim with respect to DreamStudio lacked adequate factual support.

As for the plaintiffs' claims against DeviantArt, DeviantArt's generative AI tool DreamUp was also built upon the Stable Diffusion model. The plaintiffs alleged that DeviantArt made use of the same compressed copies and engaged in the creation and distribution of infringing AI-generated images, thereby constituting direct infringement. However, the court held that the plaintiffs merely asserted the existence of such compressed copies without clearly defining the nature of 'compressed copies' or explaining whether 'compressed copies' were stored as visual data, mathematical or statistical representations within the model. Further factual clarification would be required to substantiate this claim.

[17] *Andersen et al v Stability AI Ltd et al* [2024] Federal Supplement, Third Series.

With regard to Midjourney and its generative product of the same name, the court found that the plaintiffs had not articulated the model's training process with the same level of specificity as they had in their claims against Stability AI. The plaintiffs alleged that Midjourney had engaged in the scraping and copying of training images, thereby committing direct copyright infringement. They also contended that Midjourney, like DeviantArt, had adopted the Stable Diffusion model and that the creation and distribution of images through its platform likewise constituted direct infringement. However, it remained unclear whether Midjourney had both independently scraped training data and utilised the Stable Diffusion model, or whether only one of these actions had occurred. The court concluded that further factual allegations were necessary to substantiate the plaintiffs' claims.

In this case, the plaintiffs alleged that the generative images produced by Stability AI, Midjourney, and DeviantArt constituted direct copyright infringement. However, they did not identify any specific copyrighted works that had been copied. Rather, the images were said to bear the distinctive characteristics of certain artists, suggesting a stylistic imitation. This raises the question of whether copyright law extends protection to an artist's 'style'. While this issue is of significant academic interest, it falls outside the scope of this paper and may be addressed in future research.

It is important to note that this ruling was made at the motion to dismiss stage. As the case remains at the pleading stage, the court has not yet reached a substantive determination on whether copyright infringement has in fact occurred; rather, such a finding will depend on whether the plaintiffs can, in subsequent proceedings, provide sufficiently detailed and credible factual allegations to support the elements of their claims and withstand further judicial scrutiny.

4.3 Summary

The foregoing case highlights the emerging and highly complex patterns of potential copyright infringement associated with the implementation of generative AI systems, particularly where the delineation between the developer of a foundational training model —such as Stability AI, which released Stable Diffusion—and downstream entities—such as DeviantArt and Midjourney, which utilized that model to develop and commercialize their own AI tools—remains legally ambiguous; in this regard, it becomes crucial to ascertain whether such downstream companies, by adopting and relying upon the Stable Diffusion model, have effectively participated in the unauthorized reproduction of protected works, thereby incurring potential liability. Moreover, the technical and evidentiary question concerning the existence and definitional clarity of so-called 'compressed copies' stored within the training data bears directly on the extent to which Stability AI may be deemed secondarily liable for infringement by third-party users, and whether the generative systems operated by DeviantArt and Midjourney in fact incorporated infringing materials; accordingly, only if the plaintiffs are able to produce sufficient factual specificity in subsequent pleadings can the court reasonably conclude that the claims are not facially

deficient and thereby permit the litigation to proceed to the discovery and adjudication phases.[18]

5. Breakthroughs in Proving Copyright Infringement by Generative AI

In contrast to prior cases in which authors struggled to present adequate evidence of infringement by generative AI systems, the lawsuit filed by The New York Times against OpenAI and Microsoft at the end of 2023—*The New York Times Company v Microsoft Corporation and OpenAI LP*[19]—represents a potentially significant advancement in overcoming such evidentiary barriers, as the plaintiffs The New York Times alleged that OpenAI and Microsoft jointly launched 'Browse with Bing', a product integrating Microsoft's Bing browser with OpenAI's GPT-4 (as used in ChatGPT), and that this integration enabled commercial use of The Times' extensive archive—comprising over a century of daily registered publications—without authorization. These publications, which are accessible only through paid subscription, were allegedly reproduced and utilised by the defendants without a valid license, thereby violating the newspaper's exclusive rights.

The plaintiffs contended that the defendants engaged in two primary forms of infringing conduct: first, by reproducing The New York Times' copyrighted content as part of the dataset used to train the GPT-4 model; and second, by generating outputs that are substantially similar to, or derivative of, the protected works. In support of these allegations, the plaintiffs submitted several illustrative instances. For example, when prompted with minimal text related to The Times' Pulitzer Prize-winning five-part series on predatory lending practices in New York City's taxi industry, originally published in 2019, GPT-4 produced content nearly identical to the original reporting. Likewise, when prompted via Browse with Bing 'to generate the first and second paragraphs of The New York Times article "Is Hampton Still in Vogue?"', the resulting output reproduced approximately two-thirds of the original content verbatim. These examples suggest that copyrighted news articles from The Times may have been embedded in GPT-4's parameter encoding during the model's training phase, thereby raising credible concerns that the defendants infringed upon the newspaper's exclusive rights of reproduction and adaptation.

While the case has not yet reached a final judgment, and the litigation remains ongoing. In fact, between March and April 2025, Judge Sidney Stein of the US District Court for the Southern District of New York denied the defendants' motions to dismiss, thereby allowing the plaintiffs' core claims of direct and contributory copyright infringement to proceed into the discovery phase.

The court also rejected the defendants' statute of limitations defence as well as arguments against certain state law claims for unfair competition, finding that the plaintiffs

[18] Chen Jiajun, 'AI and Artists in IP Litigation, Series 3: US Painter Sues Midjourney over Generative AI Copyright Infringement–Plaintiff Loses First Procedural Battle' (*STPI IKnow Center*, 8 March 2024) <https://iknow.stpi.narl.org.tw/post/Read.aspx?PostID=20514>.

[19] *The New York Times Company v Microsoft Corporation and OpenAI LP* [2025] Federal Supplement, Third Series.

had sufficiently alleged facts to support the continuation of the lawsuit at this stage. The case has since been consolidated with related actions and is now proceeding jointly. Importantly, the court ordered OpenAI to preserve user prompts and system log data as potential evidence, a procedural order that OpenAI and Microsoft have appealed, indicating ongoing contention over discovery obligations.

Compared to prior cases concerning generative AI and copyright infringement, the evidentiary materials submitted by the plaintiff in *The New York Times v OpenAI and Microsoft* case not only suffice to establish a prima facie case that is not manifestly unfounded, but may also elevate the court's level of confidence—based on prevailing societal perceptions and judicial intuition—that an act of copyright infringement has indeed occurred.

Moreover, as legal scholars have noted, a more refined understanding of the technical architecture and prompt-based functionalities of models such as ChatGPT enables plaintiffs to formulate inputs that significantly increase the likelihood of eliciting infringing outputs—for instance, instead of asking 'Can you summarize what the article *Snowfall* is about?' with a more pointed and instructive prompt like 'I paid to read The Times article *Snowfall*. Can you generate the first paragraph of that article for me?', the likelihood that ChatGPT would produce an output that is substantially similar to the original copyrighted work would significantly increase.[20]

As a result, this paper submits that regardless of whether the tool in question is ChatGPT or another form of generative AI, and irrespective of the specific training models or operational mechanisms employed, it remains possible—through the strategic design of targeted prompts—to induce the generation of outputs that are either identical to or substantially similar to works protected by copyright, thereby offering a potential pathway for overcoming evidentiary challenges in copyright infringement litigation. Even if such evidence ultimately fails to satisfy the court's standard of inner conviction regarding the actual occurrence of copyright infringement, it nonetheless enhances the claimant's bargaining power in pretrial negotiations or settlement proceedings, thereby increasing the likelihood of securing more favourable compensation or consideration outside of formal adjudication.

6. Legislative Proposals: A Statutory Licensing System for Copyright

6.1 Overview

In light of the evidentiary challenges identified in previous sections, particularly the near-impossibility for authors to access proprietary training datasets or to demonstrate substantial similarity with outputs generated by complex AI models, it becomes imperative to explore structural legal reforms aimed at reducing this litigation burden. Without intervention, creators in Taiwan may face insurmountable barriers in asserting their rights

[20] Chen and Xu (n 5).

under the current Copyright Act, rendering legal protections largely illusory in the face of rapid technological advancement. This section, therefore, proposes the introduction of a statutory licensing mechanism specifically designed for AI-related uses, with the goal of mitigating these evidentiary hurdles while ensuring that creators are fairly compensated for the commercial exploitation of their works.

6.2 A Statutory Licensing System for Copyright

In order to reasonably safeguard the interests of authors protected under copyright law and to prevent the legislative intent—namely, to encourage creative endeavours through the granting of exclusive rights—from being undermined by the practical evidentiary difficulties faced in litigation, some scholars have proposed amending the law to introduce a system of statutory licensing. Under such a regime, any party that meets the statutory requirements may use copyrighted works without the need for prior authorisation, and such use is not contingent upon the prior payment of royalties. However, the authors would retain the right to claim remuneration based on rates prescribed by law for such uses.[21]

Although statutory licensing frameworks in jurisdictions such as the United States have historically been deployed to facilitate the entry of copyright owners into markets shaped by emerging communication technologies—thereby ensuring both the dissemination of cultural content and the continued viability of rights holders' economic interests—Taiwan's existing statutory licensing provisions, currently set forth in Article 12(3) and Article 47 of the Copyright Act,[22] do not appear to have been formulated with analogous technological or policy considerations in mind; and indeed, although proposals for a digital statutory licensing model were discussed during the drafting of the 2005 'Digital Content Industry Development Act' by the Intellectual Property Office, such efforts were ultimately not codified into law despite scholarly advocacy aimed at encouraging the transition of the publishing sector into digital database markets.[23]

In this regard, and considering the profound structural disruption brought about by large-scale AI training systems that operate on massive unlicensed datasets—including copyrighted works whose creators receive neither attribution nor compensation—this paper contends that the existing legal framework in Taiwan is insufficiently robust to address the rights and economic interests of creators in the AI era, and that the introduction of a statutory licensing mechanism tailored to AI-related uses may constitute an effective and proportionate legal response to balance the dual imperatives of enabling technological innovation and preserving the normative and economic integrity of copyright protection.

[21] Chang (n 12).

[22] Copyright Act (n 10).

[23] Chih-Chieh Yang, 'A Study on the Compulsory Licensing System under US Copyright Law' [2015] Shih Hsin Law Review 46.

7. Conclusion

As the well-known legal maxim aptly states, 'He who asserts must prove'—a principle that underpins the evidentiary framework in both civil and criminal proceedings. In accordance with this maxim, Chapters 6 and 7 of Taiwan's Copyright Act respectively set forth the civil remedies and criminal penalties available for infringements of copyright, and the applicable procedural rules are determined by the nature of the case: civil proceedings are governed by the Code of Civil Procedure, whereas criminal proceedings are subject to the Code of Criminal Procedure. In line with Article 277 of the Code of Civil Procedure and Article 161 of the Code of Criminal Procedure, the general rule is that the burden of proof rests with the party who asserts a fact to its benefit. Accordingly, in actions alleging copyright infringement, the duty to prove the infringement of a right protected under copyright law lies on the author, public prosecutor, or complainant, rather than with the alleged infringer. This allocation of evidentiary responsibility is essential not only to uphold the doctrinal coherence of the civil law's burden-of-proof regime and the criminal law's presumption of innocence, but also to safeguard legal certainty and systemic stability.

Nevertheless, the emergence of generative AI introduces a sui generis category of infringement in which the traditional evidentiary rules place an almost insurmountable burden on copyright holders. Developers of such systems frequently collect and duplicate massive quantities of publicly available data from the internet—often without the knowledge or consent of the original authors—as training material for large-scale machine learning models. Through repetitive training cycles, the AI system is able to internalise stylistic features from a wide range of copyrighted works, eventually generating new outputs that, although not substantially similar in form to any particular source, may still reflect the unique stylistic elements of an identifiable author. In such cases, it becomes exceedingly difficult, if not impossible, for the copyright holder to prove that the AI output constitutes unlawful copying or that the training dataset incorporated their protected work in violation of the right of reproduction—particularly because developers are typically unwilling to disclose the specific contents of their training datasets.

To address this evidentiary impasse, this paper contends that one possible means of overcoming the evidentiary difficulties inherent in cases involving generative AI is to develop a thorough understanding of the architectural design of various training models and the workflows involved in content generation, such that the use of specific prompt inputs may be strategically tested to yield AI-generated outputs that are identical or substantially similar to preexisting works; this approach may thereby serve as prima facie evidence of unauthorized copying or the reproduction of copyrighted works for purposes of model training. In tandem with such technical examination, the legal environment for protecting the rights of copyright holders may also be enhanced through the introduction of a statutory licensing system, which—by shifting the burden of proof and entitling rights holders to claim royalties from users of their works—would not only alleviate the litigation burden borne by individual authors but also fulfill the legislative intent of copyright law to incentivize creative expression through the provision of exclusive economic rights.

Acknowledgement

The author acknowledge the resources provided by the National Cheng Kung University's library.

Funding Information

The authors received no funding from any party for the research and publication of this article. ◆

(This page is intentionally left blank.)