DATASET TRANSPARENCY AND COPYRIGHT ACCOUNTABILITY IN AI TRAINING SYSTEMS

Ruchi Baid*

ABSTRACT

AI companies are reluctant to reveal their training datasets, arguing that it could expose trade secrets, compromise proprietary algorithms, or reveal sensitive personal data. However, without knowing what is being fed to the algorithms, it becomes nearly impossible for the rightsholder or artist to prove that their work has been used without their permission. The case of Hayao Miyazaki is a prime example. Despite his public opposition to AI-generated Ghibli-style art, there was no easy way for him to verify whether his works were used in training models, leaving him powerless to enforce his rights. This paper argues that when artists cannot know whether their works have been used to train AI systems, copyright protection becomes illusory rather than real. Further, the onus of proof for copyright infringement allegations requires that plaintiffs prove unlawful use and sufficient similarity between the work and the copyrighted work being copied. In the context of AI technology, establishing the burden of proof becomes exceedingly difficult, if not impossible, when the underlying data utilized lacks transparency and remains undisclosed. This state of affairs also brings about asymmetry because artists cannot prove their rights without knowledge about the data used by developers of AI. The issue is further complicated by the nature of generative AI models, which do not store or reproduce specific images but learn patterns and styles from large datasets. As a result, proving the case of direct copying is significantly problematic since the outputs are not literal copies but synthetic generations based on learned features. The paper proposes policy reforms and provenance tools (Whitebox AI, watermarking, cryptographic hashing, blockchain tracking) to verify usage, enforce rights, and require dataset disclosure. The absence of these tools and a legal disclosure mandate currently permits unauthorized exploitation of artists' work, thus weakening copyright protection.

* Student, Jindal Global Law School, India

KEYWORDS: Generative Artificial Intelligence, Copyright Infringement, AI Training Data, Algorithmic Accountability, Fair Use and Fair Dealing, Protection of Creative Labour.

INTRODUCTION

Copyright law has changed a lot over time because of technology like the printing press and cameras, and now the internet. Each time something new came along, it made people question the rules about copyright, but nothing has shaken the very basics of copyright law as much, as Generative Artificial Intelligence (AI), which can create things on its own. Generative AI is different from other technologies. It does not just copy creative things. It actually learns from them. This happens when the AI looks at a lot of text, images, films, and music. The AI system starts to understand the patterns that make these things expressive, then it creates work similar to what artists create, and all this happens inside the AI algorithm system without the process being visible to the general public or to artists.

At the centre of this transformation lies a critical but unresolved question: Who is training these models, and whose creative labour is being used in the process?

This paper argues that the refusal of AI developers to disclose training datasets has produced a structural collapse of copyright enforcement. When artists cannot know whether their works have been used, they cannot assert infringement. Copyright protection in such circumstances just becomes illusory.

THE PROBLEM OF DATASET OPACITY IN GENERATIVE AI

Dataset opacity[229] indicates the lack of transparency when it comes to provenience, composition and use of databases, used to train artificial intelligence engines. Opacity in the generative AI environment is present when the developers do not make public information about the inclusion of copyrighted literary, artistic, or audiovisual works in training corpora. This lack of disclosure has manifested itself as one of the most dominant structural problems in modern copyright jurisprudence.

Generative AI systems are based on large collections of corpora in the form of textual, visual, musical, and audiovisual data collected via web scraping, licensed repositories, or amalgamated public databases. During the learning stage, these corpora are transformed into numbered vectors, which help models to identify style patterns, linguistic forms, and

[229] Adam Buick. Copyright and AI training data transparency to the rescue? Journal of Intellectual Property Law & Practice. Volume 20 Issue 3. March 2025. Pages 182–192. https://doi.org/10.1093/jiplp/jpae102

composition techniques. Although pattern recognition is a common term practitioners use to describe this operation, jurisprudence and scholarship are increasingly recognizing the reproduction of copyrighted works to be part of machine learning training, as well as the computational analysis of such works. As Lemley and Casey[230] argue, machine learning can be considered as the copying on an unprecedented magnitude, where the copyrighted media informs the functioning structure of the AI systems as an attempt.

Opacity of datasets has serious legal implications. The developers often conceal disclosures about training datasets, providing reasons of safeguarding trade secrets, design of algorithms, and positioning. Even though such arguments might be economically justified, they create an unequal structure between the creators of AI and the owners of copyright. The developers have the exclusive information on dataset composition, and the creators do not have any facility to determine whether their creations are being used in training procedures. Such an imbalance will reduce the enforceability of copyright protections.

Traditional definitions of the infringement of copyright require that the plaintiffs prove that they had access to the original work and that there is some material similarity between the alleged work that infringed and the work under protection. In the situation of the dataset opacities, though, access is established almost impossible. The intermediate or ephemeral copying has long been recognized as infringement by the courts, as was the case in the case of MAI Systems Corp. v. Peak Computer, Inc,[231] where the Ninth Circuit determined that loading computer software into a computer memory created an infringement of copyright. Similarly, these steps of digesting and training the AI might also constitute reproduction of copyrighted works. However, where no disclosure of data is made, plaintiffs do not have the evidentiary test necessary to initiate such actions.

This case directly supports the central argument advanced in this paper: copyright harm in generative AI systems often arises at the input stage rather than at the output stage. Traditionally, infringement analysis has focused on whether the final product reproduces protected expression.

THE BURDEN OF PROOF IN AI COPYRIGHT INFRINGEMENT

Within the copyright law, even an act of a simple demonstration of stylistic similarity is not considered infringement. Copyright is a safeguard of the original expression of an author and

[230] Lemley, Mark A. and Casey, Bryan, Fair Learning (January 30, 2020). Available at SSRN: http://dx.doi.org/10.2139/ssrn.3528447

[231] MAI Systems Corp. v. Peak Computer, Inc., 991 F.2d 511 (9th Cir. 1993).

not abstract ideas or artistic style. The distinctive aesthetic of Studio Ghibli[232] cannot be protected, despite individual Ghibli works otherwise qualifying for protection under copyright. To win the case concerning the infringement, a plaintiff needs to ordinarily prove (a) that the defendant accessed the copyrighted work, and (b) that the supposedly infringing production is substantially similar in the expression that it is actually protected. The first one is usually met by demonstrating that the accused had access to the original. The necessity of copying, in its turn, presupposes proving that the content that was secured was in fact used.

With regard to generative forms of artificial intelligence, such requirements are specifically hard to meet. Plaintiffs need to demonstrate that the plaintiff's work was included in the model training data. A ruling by the courts has said that copying evidence has to go beyond stylistic similarity. For example, in Getty Images v. Stability AI,[233] the court noted that a claimant has to prove that works have been (i) downloaded; (ii) by the defendant and (iii) within the boundaries of the jurisdiction of the United Kingdom. In the case of AI companies refusing to disclose their training datasets, the plaintiffs can often not meet these conditions.

Once a copying has been proved, then the plaintiff needs to demonstrate substantial components of the plaintiff's work in the output of the AI. Models produce new images or text that can only partially coincide with the inputs. The UK court admitted that models do not hold a copy of the photographs, thus ruling out the possibility of the model being a literal reproduction of the input. This fact undermines the arguments of verbatim copying. As a result, the courts decide based on the presence of enough elements of the expression that has been protected by the author in the output to qualify as infringement.

The plaintiff has a foregone burden of providing the burden of proof. According to the United States and other courts, plaintiffs are expected to prove real harm or loss (i.e., lost revenue or reputation damage). In Raw Story Media v. Open AI,[234] the case was dismissed after the defendant was unable to determine a particular article replicated by the language model and a resultant loss of income. The TechPolicy Analysis inferred that to achieve success in a court, there has to be evidence of material harm, like when an AI result or output is used to weaken a paywall or replace a business opportunity. In line with this, the plaintiffs will have to prove that their works were actually utilized and also had a significant negative effect on their market.

---

[232] Studio Ghibli vs AI: tribute or copyright infringement? (2025, April 15). IP Helpdesk. European Commission https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/studio-ghibli-vs-ai-tribute-or-copyright-infringement-2025-04-15_en#:~:text=From%20a%20legal%20point%20of,there%20could%20be%20an%20infringement

[233] Getty Images Inc. & Ors v. Stability AI Ltd. [2025] EWHC 2863 (the Ch) (Nov. 4, 2025).

[234] Raw Story Media, Inc. et al v. OpenAI Inc. et al, 1:24-cv-01514 (S.D.N.Y. Feb. 28, 2024).

These standards of evidence represent a virtually impossible obstacle in case training data are considered to be confidential. Without transparency, an artist cannot even determine whether an AI has been trained on their work, not to mention correlating a certain output with a particular work that he or she produced. Such a lack of access to the underlying data typically prevents the plaintiffs from providing tangible evidence. In line with this, courts will hardly incur to copy-cut stylized outputs.

EVOLVING JUDICIAL APPROACHES TO COPYRIGHT LIABILITY IN GENERATIVE AI

However, now courts are beginning to consider whether the reproduction of copyrighted works during training itself may be legally actionable. This changing view highlights a legal observation that infringement that can be ascribed to generative AI can occur at the time of ingesting the data and not solely by the final generated product.

One of the prominent trends is the lawsuits in the case of Andersen v. Stability AI,[235] in which a group of visual artists have filed proceedings against generative AI image-generation systems developers under the copyright infringement statute. Many of the past cases where AI came into conflict with copyright were often dismissed during the initial phase, due to difficulty in proving that there was overt copying. However, in this case, the court has taken a more liberal approach and allowed core copyright claims to proceed. By doing so, the court noted the rights of the model theory, according to which the AI models can contain the transformed iterations of the copyrighted material used in the training. Moreover, the court also acknowledged the distribution theory that argues that the spreading of AI-powered systems that are to be used with reference to the copyrighted content can constitute the distribution of infringing content. By permitting the case to move into the discovery stage, the court emphasised that liability may depend on whether protected works are embedded within AI systems in some identifiable or transformed form. This decision represents a significant doctrinal shift, as it acknowledges the possibility that training-stage use of copyrighted works may fall within the scope of infringement analysis.

Accompanying an identical trend is a parallel development of Indian law through a similar approach as it is being experienced in the case of ANI Media Pvt. Ltd. v. OpenAI.[236] The case is the first legal case in India to involve a legal inquiry into whether the use of copyrighted material in training large language models might be considered a fair dealing under Section 52

[235] Andersen, S., et al. v. Stability AI Ltd.. et al. No. 3:23-cv-00201.(N.D. Cal. filed Jan. 13, 2023).
[236] Asian News International (ANI) v. OpenAI Inc. & Anr. CS (COMM) 1028/2024. (Delhi High Court 2024).

of the Copyright Act, 1957.[237] The case foregrounds the basic questions of whether storage or processing or analysis of copyrighted news content in the context of AI training is considered as reproduction, under the Indian copyright law. More importantly, the court case places into focus the lack of express statutory safeguards in terms of the text-and-data mining or AI training within the Indian copyright paradigm. Thus, the case has triggered broader policy reflection and regulatory review, leading to the appointment of expert committees to review possible legislative change. The ANI conflict, therefore, forms a key element in a move toward technological innovation in response to copyright law in new jurisdictions.

COMPARATIVE GLOBAL REGULATORY APPROACHES

With the rapid growth of generative artificial intelligence, different countries have attempted to enact laws in line with their respective priorities in terms of balancing between innovation and copyright protection. The doctrine of fair use is most common in the United States. Courts evaluate whether the use of copyrighted material in an incorporation has become a sufficiently transformative use and whether it will cause harm to the market of the original work. Recent court cases, e.g. Thomson Reuters v. ROSS Intelligence,[238] portray that the artificial intelligence training can be beyond the fair-use protection when the same is used commercially or in competition with the market of the creative work of the original content. Other U.S. cases have, in contrast, confirmed the transformative quality of AI training whereby resultant outputs lack the re-creation of expressive content or something that causes a quantifiable economic harm, which underscores the fact-driven and adaptive nature of American copyright case law.

In contrast, the European Union has taken a more organized regulatory strategy, being implemented in the EU Artificial Intelligence Act[239] and text-and-data mining of the Copyright Directive. Such frameworks clearly acknowledge the possibility of involvement of reproduction rights in the training of AI, and require developers to address transparency principles.

India stands halfway, providing a strong assurance of the protection of the expressive works, but not providing a clear statute concerning AI training at the given moment. This uncertainty[240] is echoed in the constantly pending courts of law, legal cases, and policy

[237] The Copyright Act, 1957. No. 14.Acts of Parliament.1957.
[238] Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc.No. 1:20-cv-00613-SB. (D. Del. Feb. 11, 2025).
[239] Article 50: Transparency obligations for providers and deployers of certain AI systems | EU Artificial Intelligence Act. (n.d.). https://artificialintelligenceact.eu/article/50/
[240] AI & copyright.Udit Mendiratta.Argus Partners Blog.26th Mar, 2024. https://www.argus-p.com/papers-publications/thought-paper/ai-copy,,,right/

discussions. The differential regulatory methods that are brought into focus by these jurisdictions highlight the lack of a coherent standard due to differences in each jurisdiction‘s priorities and frameworks to balance technological progress and copyright protection.

RECOMMENDATIONS

An important reform involves the evidentiary burden of litigation involving works created through artificial intelligence (AI) and copyright matters. The current legal provisions put the whole burden of the proof on the copyright holders, and they are required to prove access, reproduction, and similarity. Nonetheless, the vagueness of the training data hinders the ability of creators to acquire the evidence that meets these needs. As a result, this asymmetry establishes structural barriers to enforcing it and weakens the strength of the copyright protection.

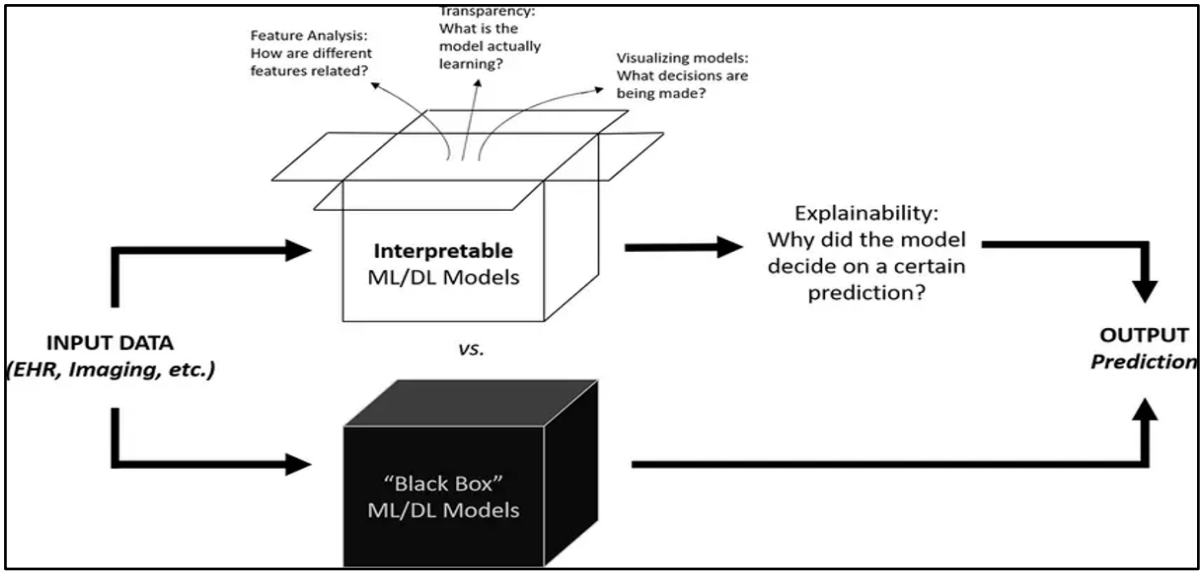
To address this issue, jurisdictions can consider the adoption of rebuttable presumptions in case of AI copyright cases. In this case, the evidence burden may move to the AI developers in a situation where owners of the copyright display a reasonable level of similarity between the works and the outputs created by AI. Such an approach does not eliminate the evidentiary standards, it only readjusts enforcement instruments, accepting the fact that developers have better access to the relevant information.

TECHNICAL SOLUTIONS WHITE-BOX AI AND PROVENANCE MECHANISMS

They should be supplemented by technological interventions that can initiate change in the context of increasing transparency and traceability in the process of AI training. The shift to more transparent or white box AI models, as opposed to opaque or black box models, is an important step towards accountability.

The diagram below[241] compares interpretable models, which can be analysed in terms of features, visualise decision-paths, and provide traceability between inputs and outputs, with black-box models, which are represented by most generative AI solutions and whose outputs are not accompanied by an explanation of internal computations. This lack of explainability in itself contributes to the lack of explainability in datasets and asymmetry of evidence. The developers can achieve the goal of providing auditability by adopting explainable AI frameworks without reducing the technological efficacy.

[241] Hui, Aaron & Ahn, Shawn & Lye, Carolyn & Deng, Jun. (2022). Ethical Challenges of Artificial Intelligence in Health Care: A Narrative Review. Ethics in Biology, Engineering and Medicine: An International Journal. 12. 10.1615/EthicsBiologyEngMed.2022041580.



Source: Research Gate (Ethical Challenges of AI in Health Care, Hui et al.)

White-box artificial intelligence requires[242] interpretable models, such as decision trees, linear regression, and rule-based algorithms, which allow the stakeholders to determine the effect of the input variables on the predictions of the output. This obligatory disclosure makes white-box AI part and parcel of systems of accountability like watermarking, cryptography hashing, and provenance tracing. Since the paths of decision-making of such models are clearly visible, the regulators and the rights holders can easily confirm the influence of the copyrighted works on the output of AI. Hence, white-box AI justifies transparency and eliminates information asymmetry between developers and creators.

White-box AI models, nevertheless, also have serious weaknesses. They work well typically in low- to medium-complexity data climates and are incapable of processing intensely complicated or massive data volumes. In comparison, more developed systems of generative AI (large language models and image diffusion models) make use of black-box neural networks that can analyse large, unstructured data. These models allow an improved predictive accuracy, creative synthesis and scaling, which makes them essential to the modern AI development.

Although black-box AI provides better performance, its inability to produce any form of explanation raises their issues with transparency, as well as copyright protection. In spite of lack of internal transparency in black-box systems, copyright protection can still be achieved by the combination of regulatory disclosure requirements and technological responsibility tools. Seeing that black-box models cannot entirely reveal their learning procedures,

[242] Ihejirika, C. J. (2025). Explainable artificial intelligence in deep learning models for transparent decision-making in high-stakes applications. International Journal of Research Publication and Reviews, 6(2).

governance should be directed to their training-data consumption, traceability of their results, as well as compliance by developers, and not necessarily to the examination of their pathways to arrive at the final decision.

One of the main ways of protecting copyright in black-box AI is related to the disclosure of datasets. Laws and regulations, including the European Union AI Act,[243] require developers to give descriptions or documentation of training datasets used in AI creation. Though not requiring the disclosure of proprietary algorithms, exhaustive data sets etc., this allows the copyright owners to ascertain whether their work has been used in training regimes or not. These transparency provisions reduce the unilateral nature of information, and they enable the rights holders to challenge unwarranted use.

The other course of action is to have training data audit measures. In these systems, training datasets of AI can be reviewed by separate regulatory organizations or accredited auditors under privacy protection. It is a method that allows checking that the copyright has not been violated without compelling the developers to publicly expose commercially sensitive information. A balance between trade-secret security requirements and the requirements of accountability is achieved through confidential audit models.

Efficient protection of copyright in black-box AI can also be provided through technological solutions. Digital watermarking[244] allows authors to add imperceptible identifiers to creative content, allowing them to track their usage by the training datasets or AI results. Similarly, cryptographic hashing gives copyrighted material cryptographic fingerprints that will permit content creators or authorities to identify whether a dataset has been added to a dataset without disclosing proprietary datasets. There are also provenance-tracking systems based on blockchains, which provide secure and irrevocable records that provide a connection between copyrighted content and licensing agreements, as well as the usage of datasets, increasing evidential bases in copyright litigation.

Finally, black-box AI systems require copyright protection, which requires a change in enforcement methods to data-governance transparency, as opposed to algorithmic transparency.[245] Regulatory frameworks are preferring accountability, as demonstrated by

[243] Article 50 & 51: Transparency obligations for providers and deployers of certain AI systems | EU Artificial Intelligence Act. (n.d.).

[244] Gupta, R., EY India, Vij, J., FICCI, & Singh, A. (2024). Identifying AI generated content in the digital age: The role of watermarking [Report].

[245] Rico, Philippe. (2024). AI and Data Governance: A Legal Framework for Algorithmic Accountability and Human Rights.

documentation of datasets, auditability, licensing checking, and provenance tracing, by avoiding the need to disclose the inner workings of the neural network.

CONCLUSION

While regulation is necessary to address copyright challenges in generative AI, excessive or rigid regulation may slow technological innovation by restricting access to training data and increasing compliance costs. Such constraints may discourage research investment and fragment global AI development. However, the absence of regulation also poses serious risks, as generative AI often relies on large-scale use of copyrighted works, potentially enabling the extraction of creative labour without consent or compensation. The challenge, therefore, lies in achieving regulatory balance. Effective frameworks must allow AI technologies to grow while ensuring that creators retain meaningful protection over their works. Transparency measures, licensing mechanisms, and accountability tools can help maintain this balance. Ultimately, sustainable AI development requires governance models that support technological progress while protecting the creative labour from being exploited by artificial intelligence systems.