

Article

Generative AI, Copyright Law, and the Future of Creativity: A Call for International Reform

Aminesh Nayan Gogoi*

India International University of Legal Education and Research, Goa, India

* Correspondence: amineshnayangogoi@gmail.com

Abstract: Current study examines whether the use of copyrighted works in training generative AI systems amounts to copyright infringement under existing legal frameworks. The challenge is not merely to fit AI within the existing contours of copyright law, but to rethink those contours in light of new technological realities. The goal should be to construct a legal framework that preserves the foundational objectives of copyright, encouraging creativity and protecting authors, while enabling the continued growth of transformative technologies. In this sense, the regulation of AI training represents not an endpoint, but an evolving legal frontier that will shape the future relationship between law, creativity, and innovation. Current research analyzes the applicability of doctrines such as fair use and fair dealing in the context of large-scale data scraping and machine learning processes. It explores challenges relating to transparency, attribution, and enforcement, given the opaque nature of AI training datasets. By engaging with emerging global legal developments, it assesses whether current copyright regimes are equipped to address AI-driven innovation, and argues for clearer standards to balance technological advancement with the protection of creators' rights in the digital environment.

Keywords: artificial intelligence; copyright law; generative AI; fair use; fair dealing; intellectual property

1. Introduction

The emergence of generative artificial intelligence (AI) marks a significant shift in the technological and creative landscape. Unlike earlier computational systems that relied on explicitly coded instructions, generative AI models, such as large language models and image generation systems, operate on data-driven architectures built through machine learning, particularly deep learning techniques. These systems are trained on vast datasets to identify patterns and generate outputs that resemble human-created content (Goodfellow, Bengio, and Courville 2016). The effectiveness of such systems is directly dependent on the scale, diversity, and quality of the data used during training, making data a central resource in the development of modern AI technologies.

This data-driven paradigm, however, has brought generative AI into direct conflict with copyright law. On one hand, innovation in AI requires broad access to data to enable meaningful learning and competitive development. On the other hand, a substantial portion of the data used in training consists of copyrighted works, including literary, artistic, and audiovisual content, often utilized without explicit authorization from rights holders (European Parliament 2020). This creates a structural tension between fostering technological progress and protecting the economic and moral rights of creators, particularly where AI systems may derive value from such works without compensation or attribution.

Against this backdrop, the central research problem of this paper is whether the use of copyrighted works in AI training constitutes an infringing use under copyright law. The issue is complex, as AI training involves acts of copying and processing data that may fall within the scope of exclusive rights such as reproduction, yet may also be argued to constitute non-expressive or functional use. This research examines the legality of using copyrighted works in the training of generative artificial intelligence (AI) systems, with a primary focus on the processes of data ingestion, copying, and analysis during the training phase. It is situated within the framework of copyright law and adopts a comparative approach, referring to key jurisdictions such as the United States, the United Kingdom, and India. The study aims to analyze whether AI training falls within the scope of exclusive rights, particularly reproduction, and whether it can be justified under exceptions such as fair use or fair dealing. It further seeks to identify regulatory and enforcement challenges, including issues of transparency, attribution, and jurisdiction, while evaluating emerging global legal approaches. Ultimately, the objective is to propose balanced legal solutions that reconcile the demands of technological innovation with the protection of creators' rights.

Citation: Aminesh Nayan Gogoi. 2026. Generative AI, Copyright Law, and the Future of Creativity: A Call for International Reform. *Trends in Intellectual Property Research* 4(3), 21-26. <https://doi.org/10.69971/tipr.4.3.2026.136>



Copyright: © 2026 by the authors. This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>.

The study *AI, Authorship, and Copyright: A Comparison Between the United States and the European Union* (2025) systematically deconstructs how both the United States and the European Union currently handle the authorship of AI-generated works, highlighting a shared judicial reluctance to expand the definition of a "creator" beyond a human being. In the U.S. context, the journal meticulously dissects the fallout from the *Thaler* decision and the U.S. Copyright Office's subsequent guidance, which strictly mandates that copyright protection only attaches where a human contributes "original expression." The analysis of the EU approach is similarly grounded, focusing on the *Standard Verlags GmbH* standard, which dictates that a work must reflect the author's personality, a criterion inherently impossible for an autonomous AI to fulfill ("*AI, Authorship, and Copyright*" 2025).

Craig and Kerr rely on relational copyright theory to argue forcefully against the assignment of authorship to artificial intelligence. They posit that authorship is inherently an act of human-to-human communication, a culturally embedded dialogue that a machine, lacking consciousness, lived experience, and socialization, is in its nature, incapable of participating in. This perspective systematically dismantles utilitarian rationalizations for machine authorship, reinforcing the argument that an AI cannot be an author because it does not communicate meaning; it merely translates mathematical weights. Consequently, the communicative act, and the resulting legal authorship, must logically vest in the human designer who conceptually engineered and curated the output for human consumption (Craig and Kerr 2021).

Samuelson (2020) addresses the pragmatic legal complexities of allocating rights when generative systems are utilized in the creative process. She explores the triangular tension between the AI programmer who develops the algorithmic architecture, the original human creators whose works constitute the scraped training archive, and the end-user who engineers the prompt. Samuelson systematically questions who, if anyone, possesses the strongest legitimate claim to the synthetic output (Samuelson 2020).

2. Methods

In current research, the doctrinal method has been used where it will be concerned with the documented research, as the library was the only source of information put to use. Internet source also provided substantive information. Secondary sources of data have been used where documents have gone through some interpretations and have been used further for the study of research problem.

3. AI Training and Use of Copyrighted Data

Generative artificial intelligence (AI) models are developed through sophisticated machine learning processes, most notably deep learning, which relies on multi-layered neural networks to process and interpret large volumes of data. The training of such models involves feeding them extensive datasets and allowing them to progressively adjust internal parameters through optimization techniques such as gradient descent. Over repeated cycles, the model learns statistical relationships between inputs and outputs, enabling it to generate text, images, or other forms of content that closely resemble human-created works. Importantly, this learning is not based on conscious understanding but on probabilistic pattern recognition derived from exposure to data at scale.

The effectiveness of generative AI systems is fundamentally dependent on the size and diversity of their training datasets. In practice, these datasets are assembled through methods such as web scraping, automated data extraction, and large-scale ingestion of publicly accessible information. Sources may include websites, online publications, image repositories, forums, and social media platforms. Given the scale and automation involved, the datasets are often compiled without granular filtering for copyright status, resulting in the inclusion of vast amounts of protected content. This process reflects a broader industry practice where the emphasis is placed on maximizing data volume to improve model performance, rather than ensuring legal compliance at the level of individual works.

Within this framework, copyrighted works play a crucial and often indispensable role in model development. High-quality expressive materials, such as literary texts, journalistic writing, visual art, music, and audiovisual content, provide the structural richness and diversity necessary for effective training (Lemley and Casey 2021). These works enable AI systems to learn grammar, narrative structure, stylistic nuances, visual composition, and cultural context. As a result, the performance and sophistication of generative AI models are closely tied to their access to copyrighted content, making such material a foundational component rather than a peripheral input.

From a legal standpoint, a critical distinction must be drawn between input (training data) and output (generated content). The input stage involves the ingestion, storage, and processing of existing works, which may implicate the exclusive right of reproduction under copyright law. During this phase, multiple copies of the data may be created, whether temporarily in memory or as part of pre-processing pipelines (*MAI Systems Corp. v. Peak Computer* 1993). The output stage, by contrast, involves the generation of new content based on learned statistical patterns. While outputs are typically novel and not direct reproductions, they may still raise legal concerns if they are substantially similar to protected works or replicate identifiable elements of style.

This distinction is central to the broader legal debate. While public attention often focuses on whether AI-generated outputs infringe copyright, the more immediate and systemic issue lies in the training process itself, where large-scale copying occurs as a necessary technical step. The challenge is further compounded by the opaque nature of AI systems. Developers frequently treat training datasets as proprietary information, limiting public access and scrutiny (WIPO 2020). This lack of transparency makes it difficult to determine whether specific copyrighted works have been used, to assess the extent of their influence, or to establish a causal link between training data and generated outputs.

Consequently, understanding AI training requires not only a technical appreciation of how models function but also a legal analysis of how existing copyright doctrines apply to these processes. The interplay between large-scale data ingestion, reliance on copyrighted works, and the distinction between input and output forms the foundation for evaluating whether AI training constitutes permissible use or actionable infringement.

4. Copyright Law and the Question of Infringement

Copyright law is premised on granting authors a bundle of exclusive rights over their original works, including the rights of reproduction, adaptation, distribution, and communication to the public.¹ Among these, the right of reproduction is particularly central to the debate on artificial intelligence (AI) training. Reproduction is broadly defined across jurisdictions to include the making of copies in any material form, including digital storage.² Similarly, the right of adaptation protects against the transformation of a work into another form, capturing derivative uses that build upon the original expression. These foundational principles establish a legal framework within which the acts involved in AI training must be assessed.

The process of training AI models raises the question of whether such activities amount to “use” or “copying” in the copyright sense. Technically, training involves the ingestion, storage, and processing of data, often requiring multiple instances of copying, whether temporary or permanent. From a strict doctrinal perspective, such acts may fall within the scope of reproduction. However, proponents of AI development argue that training constitutes a non-expressive or functional use of works, where the system does not exploit the expressive value of the content but instead extracts statistical patterns. This distinction between expressive copying and functional analysis lies at the heart of the legal ambiguity surrounding AI training.

The debate is further complicated by the question of human versus machine use. Traditional copyright infringement is grounded in human agency, requiring an identifiable actor who engages in unauthorized use of a protected work. In the context of AI, the copying is performed by automated systems operating at scale, often without direct human intervention at each instance of data processing. This raises questions about attribution of liability, whether it should attach to developers who design the models, entities that compile training datasets, or users who deploy the systems (Abbott 2020). The diffusion of responsibility challenges conventional notions of authorship and infringement, requiring a re-examination of how liability is assigned in technologically mediated environments.

Another critical aspect is the threshold of infringement in digital environments, which has evolved through judicial interpretation. Courts have recognized that even temporary or transient copies, such as those stored in a computer’s memory, may constitute reproduction if they are sufficiently fixed and integral to a technological process (*Infopaq International* 2009). At the same time, judicial decisions have demonstrated a degree of flexibility in cases involving technological innovation. For instance, in cases concerning search engines and digital indexing, courts have permitted certain forms of intermediate copying on the grounds of functionality and public benefit (*Authors Guild v. Google* 2015). These precedents suggest that not all acts of copying are treated equally, and the context and purpose of use play a significant role in determining infringement.

In the context of AI training, these principles create a complex legal landscape. On one hand, the technical processes involved appear to satisfy the requirements of reproduction. On the other hand, the functional and non-expressive nature of training, combined with its role in enabling innovation, complicates a straightforward finding of infringement. The absence of direct judicial authority on AI-specific scenarios further underscores the uncertainty, making this an area where doctrinal interpretation and policy considerations must evolve in tandem.

5. Fair Use, Fair Dealing, and Transformative Use

The legality of using copyrighted works in artificial intelligence (AI) training is often examined through the lens of exceptions and limitations to copyright, particularly the doctrines of fair use and fair dealing. These doctrines serve as balancing mechanisms, allowing certain unauthorized uses of protected works where they advance broader public interests such as innovation, research, and access to knowledge. Their applicability to AI training remains one of the most contested issues in contemporary copyright law.

Under the United States framework, the doctrine of fair use, codified in § 107 of the U.S. Copyright Act, provides a flexible, case-by-case assessment based on four factors: the purpose and character of the use, the nature of the copyrighted work, the amount and substantiality of the portion used, and the effect on the potential market.³ Courts have increasingly emphasized the concept of “transformative use,” which considers whether the new use adds a different purpose or character to the original work (*Campbell v. Acuff-Rose Music* 1994). In cases such as *Authors Guild v. Google, Inc.*, large-scale digitization and indexing of books were held to be fair use due to their transformative and non-expressive nature (*Authors Guild v. Google* 2015). Proponents of AI training argue that similar reasoning applies, as training involves extracting patterns rather than reproducing expressive content. However, critics point out that the commercial nature of many AI systems and their potential to substitute for original works may weigh against a finding of fair use.

In contrast, jurisdictions such as India and the United Kingdom follow the doctrine of fair dealing, which is more restrictive and operates through enumerated exceptions. Under the Copyright Act, 1957 (India), fair dealing is permitted for specific purposes such as private use, research, criticism, review, and reporting of current events.⁴ Similarly, the Copyright, Designs and Patents Act 1988 (UK) provides limited exceptions, including those for research and private study, as well as text and data mining (TDM) for non-commercial purposes.⁵ Unlike the open-ended nature of U.S. fair use, these provisions are purpose-specific and less adaptable to novel technological uses. As a result, the application of fair dealing to AI training, particularly for commercial models, remains legally uncertain and arguably constrained.

The concept of transformative use plays a central role in assessing AI training within the fair use framework. Transformative use requires that the secondary use does not merely replicate the original work but instead creates new meaning, expression, or utility (Leval 1990). In the context of AI, the argument is that training transforms expressive works into non-expressive data repre-

¹ Copyright Act, 1957 (India), § 14; Berne Convention for the Protection of Literary and Artistic Works, 1886, art. 9.

² Copyright, Designs and Patents Act 1988 (UK), § 17.

³ U.S. Copyright Act, 17 U.S.C. § 107.

⁴ Copyright Act, 1957 (India), § 52.

⁵ Copyright, Designs and Patents Act 1988 (UK), §§ 29–30; § 29A (text and data mining).

sentations, such as weights and parameters within a model. This transformation is functional rather than expressive, aimed at enabling the system to generate new outputs rather than to substitute the original works. However, this characterization is contested, as the transformation occurs at a computational level that may not align neatly with traditional legal interpretations of transformation.

The ultimate question, therefore, is whether AI training qualifies as a permissible use under these doctrines. While there are compelling arguments in favor of treating AI training as fair use, particularly in light of its transformative and functional aspects, there are equally significant concerns regarding market harm, unauthorized exploitation, and the scale of copying involved. In jurisdictions governed by fair dealing, the lack of flexibility further complicates the recognition of AI training as a lawful exception, especially in commercial contexts.

In the absence of definitive judicial guidance, the application of fair use, fair dealing, and transformative use to AI training remains uncertain. The issue highlights the limitations of existing doctrines in addressing technologically complex and large-scale uses of copyrighted material, underscoring the need for clearer legal standards or legislative intervention.

6. Challenges in Regulation and Enforcement

The regulation and enforcement of copyright law in the context of artificial intelligence (AI) training present significant structural and practical challenges. These challenges stem from the scale, complexity, and opacity of AI systems, which do not align easily with traditional enforcement mechanisms designed for identifiable acts of copying and distribution. As AI technologies continue to evolve, these limitations have become increasingly evident, raising concerns about the effectiveness of existing legal frameworks.

A primary challenge lies in the lack of transparency in training datasets. AI developers often treat their datasets as proprietary information, protected under trade secrecy or confidentiality claims (WIPO 2020). As a result, there is limited public disclosure regarding the sources, composition, or scope of the data used in training. This opacity makes it difficult for rights holders to determine whether their works have been included, thereby hindering the ability to assert or enforce copyright claims. Moreover, the absence of transparency complicates regulatory oversight, as authorities lack access to the information necessary to assess compliance with copyright law.

Closely related are attribution and ownership concerns. Even where it is acknowledged that copyrighted works contribute to the training process, identifying the specific authors or rights holders whose works have been used is often impractical (European Commission 2022). Generative AI models are trained on vast and heterogeneous datasets, making it nearly impossible to trace the influence of individual works on the model's outputs. This raises questions about whether and how creators should be credited or compensated, particularly when AI-generated outputs exhibit stylistic similarities to identifiable authors or artists. The lack of attribution mechanisms also undermines moral rights, which in many jurisdictions include the right to be identified as the author of a work.

Another significant issue is the difficulty in identifying infringement. Traditional copyright enforcement relies on the ability to demonstrate substantial similarity between a protected work and an allegedly infringing copy. In the context of AI, this becomes more complex, as the outputs generated by models are typically novel and not direct reproductions of any single work (Lemley and Casey 2021). Even where similarities exist, establishing a causal link between a specific copyrighted work used in training and a particular output is technically and evidentially challenging. Furthermore, the distributed and automated nature of AI training processes obscures the chain of causation, making it harder to attribute liability with certainty.

The challenges are further exacerbated by cross-border and jurisdictional issues. AI development and deployment often occur across multiple jurisdictions, with data collected in one country, models trained in another, and outputs disseminated globally (OECD 2019). This raises complex questions regarding the applicable law, jurisdiction for enforcement, and conflicts between differing national copyright regimes. For instance, a use that may qualify as fair use in the United States may not fall within the narrower fair dealing exceptions in India or the United Kingdom. The lack of harmonization in international copyright law thus creates uncertainty and inconsistency in the regulation of AI training practices.

Collectively, these challenges highlight the inadequacy of existing enforcement mechanisms in addressing the realities of AI-driven technologies. The opacity of datasets, the difficulty of attribution, the evidentiary hurdles in proving infringement, and the fragmentation of legal regimes across jurisdictions all contribute to a regulatory environment that struggles to keep pace with technological innovation. Addressing these issues will require not only doctrinal adaptation but also institutional and policy-level reforms aimed at enhancing transparency, accountability, and cross-border cooperation.

7. Recommendations

The rapid evolution of artificial intelligence (AI) has exposed significant limitations within existing copyright frameworks, particularly in their ability to address large-scale, automated, and non-expressive uses of protected works. Copyright law, as it currently stands, is largely premised on human authorship, identifiable acts of copying, and relatively discrete instances of infringement (Samuelson 2020b). These assumptions do not translate effectively to AI training, which involves continuous, large-scale ingestion and processing of data in ways that are both technologically complex and legally ambiguous. As a result, the existing framework struggles to provide clear guidance on whether such uses are permissible, leading to uncertainty for both innovators and rights holders.

One of the primary gaps lies in the absence of explicit provisions addressing text and data mining (TDM) and machine learning activities across jurisdictions. While some legal systems have begun to introduce limited exceptions, for instance, non-commercial TDM exceptions in certain jurisdictions, these remain fragmented and often insufficient to accommodate the realities of modern AI

development.⁶ Additionally, traditional doctrines such as fair use and fair dealing were not designed with algorithmic learning in mind, resulting in inconsistent and unpredictable outcomes when applied to AI-related disputes. This doctrinal mismatch underscores the need for a more tailored legal approach.

Accordingly, there is a pressing need for clearer legal standards that specifically address the use of copyrighted material in AI training. Such standards must strike a balance between enabling innovation and safeguarding the legitimate interests of creators. Clarity is essential not only for ensuring compliance but also for reducing litigation risks and fostering a stable environment for technological development (WIPO 2020). This may involve legislative intervention to define the scope of permissible uses, establish conditions under which AI training is allowed, and clarify the allocation of liability among different actors in the AI ecosystem.

Among the potential solutions, licensing regimes have emerged as a viable mechanism for reconciling competing interests. Under such frameworks, AI developers could obtain licenses to use copyrighted works in training datasets, either through collective management organizations or negotiated agreements with rights holders (European Union 2019). This approach would ensure that creators are compensated for the use of their works while providing legal certainty to developers. However, the feasibility of licensing at scale remains a challenge, given the vast number of works involved and the difficulty of identifying and negotiating with individual rights holders.

Another critical reform measure is the enhancement of data transparency. Requiring AI developers to disclose information about the sources and composition of their training datasets could significantly improve accountability and facilitate enforcement. Transparency obligations would enable rights holders to identify potential unauthorized uses and assert their rights more effectively. At the same time, such measures must be carefully designed to balance the protection of trade secrets and proprietary interests with the need for regulatory oversight.

Finally, there is growing support for the development of *sui generis* approaches tailored specifically to AI and data-driven technologies. These could include the creation of new legal categories or rights that address the unique characteristics of AI training, such as compulsory licensing schemes, opt-out mechanisms for rights holders, or statutory exceptions conditioned on compliance with specific safeguards (Guadamuz 2017). A *sui generis* regime would allow lawmakers to move beyond the constraints of traditional copyright doctrine and develop a more flexible and forward-looking framework.

In sum, the challenges posed by AI training require a recalibration of copyright law that goes beyond incremental adaptation. Addressing the identified gaps through clearer standards, innovative licensing models, enhanced transparency, and potentially new legal regimes will be essential in ensuring that the law remains responsive to technological change while preserving the core objective of protecting creative expression.

8. Conclusions

At its core, the debate reflects a deeper need to re-balance the relationship between innovation and the protection of creators' rights. On one hand, generative AI represents a powerful tool for economic growth, creativity, and knowledge dissemination, and overly restrictive legal approaches risk stifling its development. On the other hand, the unregulated use of copyrighted works undermines the incentive structure that copyright law seeks to preserve, potentially eroding both economic returns and moral recognition for creators. A sustainable legal response must therefore move beyond binary positions of strict enforcement or complete exemption and instead adopt a calibrated approach that recognizes the legitimacy of both interests. Mechanisms such as fair compensation, transparency obligations, and conditional exceptions may offer pathways toward such balance. The future of copyright in the AI era will likely depend on the ability of legal systems to evolve in response to technological change. Existing doctrines, while adaptable to some extent, are increasingly strained when applied to data-driven and automated processes at scale. This necessitates not only judicial interpretation but also proactive legislative and policy interventions. Emerging approaches, including licensing frameworks, text and data mining exceptions, and potential *sui generis* regimes, indicate a shift toward more nuanced and technology-specific solutions (European Union 2019; European Commission 2022). At the international level, greater harmonization will be essential to address the cross-border nature of AI development and deployment.

References

- Abbott, Ryan. 2020. *The Reasonable Robot: Artificial Intelligence and the Law*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781108636795>
- Craig, C. J., and Ian R. Kerr. 2021. The Death of the AI Author. *Ottawa Law Review* 52. <https://papers.ssrn.com/abstract=3374951>
- European Commission. 2022. *Study on Copyright and New Technologies: Copyright Data Management and Artificial Intelligence*. Luxembourg: Publications Office of the European Union. <https://digital-strategy.ec.europa.eu/en/library/study-copyright-and-new-technologies-copy-right-data-management-and-artificial-intelligence>
- European Parliament. 2020. *Intellectual Property Rights and Artificial Intelligence*. PE 634.452. Brussels: European Parliament. [https://www.eu-roparl.europa.eu/thinktank/en/document/IPOL_STU\(2020\)634452](https://www.eu-roparl.europa.eu/thinktank/en/document/IPOL_STU(2020)634452)
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge, MA: MIT Press. <https://doi.org/10.7551/mit-press/9780262035613.001.0001>
- Guadamuz, Andres. 2017. Artificial Intelligence and Copyright. *WIPO Magazine*. October 2017. https://www.wipo.int/wipo_magazine/en/2017/05/article_0003.html
- Landes, William M., and Richard A. Posner. 2003. *The Economic Structure of Intellectual Property Law*. Cambridge, MA: Harvard University Press. <https://doi.org/10.4159/9780674038992>
- Lemley, Mark A., and Bryan Casey. 2021. Fair Learning. *Texas Law Review* 99: 743–786. <https://papers.ssrn.com/abstract=3528447>
- Leval, Pierre N. 1990. Toward a Fair Use Standard. *Harvard Law Review* 103 (5): 1105–1136. <https://doi.org/10.2307/1341457>

⁶ Copyright, Designs and Patents Act 1988 (UK), § 29A; European Union Directive (EU) 2019/790 on Copyright in the Digital Single Market, arts. 3–4.

- OECD. 2019. *Artificial Intelligence in Society*. Paris: OECD Publishing. <https://doi.org/10.1787/2a2e2d45-en>
- Samuelson, Pamela. 2020a. AI Authorship? *Communications of the ACM* 63 (7): 20–22. <https://doi.org/10.1145/3404510>
- Samuelson, Pamela. 2020b. Implications of Artificial Intelligence for Intellectual Property Law and Policy. *Stanford Technology Law Review* 23: 1–32. <https://papers.ssrn.com/abstract=3679997>
- World Intellectual Property Organization (WIPO). 2020. *Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence*. WIPO/IP/AI/2/GE/20/1. Geneva: WIPO. https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1.pdf