

ARTIFICIAL INTELLIGENCE, DIGITAL TECHNOLOGIES & DATA GOVERNANCE

Article



DOI: 10.17803/2713-0533.2026.2.36.237-268

From Data Mining to Copyright Infringement: Legal Challenges in Training Artificial Intelligence

Siddharth Badkul

Maharashtra National Law University, Nagpur, India;

Bennett University, Greater Noida, India

© S. Badkul, 2026

Abstract: Legal controversies pertaining to the use of copyrighted material in artificial intelligence training datasets have been fuelled by the fast development of large language models. The paper explores the complex legal issues resulting from data mining activities and possible copyright violations associated with the training of AI. The paper includes the technological perspective of data mining and training of AI models. Further, the paper deals with the legal challenges, scrutinising whether specific technological methods such as the use of unrecognisable examples, watermarking techniques, machine unlearning, dataset de-duplication, etc., thereby exploring how their application prevents copyright infringement. As original contributions, the paper analyses the theories related to copyright law and evaluates the effectiveness of the aforementioned technological method in fulfilling the objectives of these theories. The paper provides an original framework of assessment of copyright issues in AI training datasets, by contrasting systems in the EU, U.S., and India. It reveals regulatory loopholes in the copyright law in India and suggests a hybridized approach to Indian copyright law, the Fair Learning Doctrine that combines the concepts of transformative

use and proportionality. The paper adopts a doctrinal and analytical approach, analysing case laws, statutory interpretation, and technological literature to assess the relationship between AI and copyright vis-à-vis training of AI model.

Keywords: text and data mining; fair use doctrine; large language models; copyright infringement; AI training data

Cite as: Badkul, S., (2026). From Data Mining to Copyright Infringement: Legal Challenges in Training Artificial Intelligence. *Kutafin Law Review*, 13(2), pp. 237–268, doi: 10.17803/2713-0533.2026.2.36.237-268

AI Use Statement: AI tools were used for grammar checking, structuring of sentences and data tabulation to enhance clarity and presentation. All intellectual content, analysis, and conclusions are the original work of the author.

Contents

I. Introduction	239
II. Data Mining and Copyright Violation: Legal Perspectives	
from European Union and United States of America	240
III. Technical Perspective	248
IV. Legal Challenges in Training AI: From Data Mining	
to Copyright Infringement	250
IV.1. Whether Infringement or not? A Jurisprudential Perspective	252
IV.2. Philosophical Approaches to Copyright and AI Training	253
IV.3. Critical Legal Perspective	254
IV.4. Suggestions: Technological methods and Strategies	256
IV.4.1. Unrecognisable Examples	256
IV.4.2. Watermarking Techniques	256
IV.4.3. Machine Unlearning	257
IV.4.4. Dataset De-duplication	257
IV.4.5. Alignment Strategies	257
IV.4.6. Improved Training Algorithms	257
IV.5. Proposed Changes in Legal Framework	259
IV.6. Proposed Integrated Framework	261
IV.7. Lessons for India	261
V. Conclusion	264
References	264

I. Introduction

The interplay between copyright law and artificial intelligence presents a multifaceted scenario, particularly with respect to the utilization of copyrighted material for training AI models. An essential element of this junction pertains to the training of these models, which frequently necessitates the utilization of extensive datasets that may encompass copyrighted materials. This method raises significant legal concerns, particularly regarding potential copyright infringement during the data-mining phase, in which algorithms extract patterns and insights from large datasets. This method raises significant legal concerns, particularly regarding potential copyright infringement during the data-mining phase, in which algorithms extract patterns and insights from large datasets. The primary concern pertains to whether the utilization of copyrighted content for training of AI qualifies as fair use or is subject to alternative exclusions within copyright law. As AI technology continues to advance, copyright considerations become increasingly critical to promoting innovation while respecting authors' rights. As AI technology continues to advance, copyright considerations become increasingly critical to promoting innovation while respecting authors' rights. The use of copyrighted photos for training generative AI systems has incited copyright infringement litigation, complicated the legal framework and underscored the conflict between technological progress and existing intellectual property rights (Quang, 2021).

The issues in this domain may be categorized into three principal stages: the learning and training phase, the generation of AI-produced content, and the operation of the AI model itself. Each stage gives rise to distinct legal challenges, particularly in relation to intellectual property rights. This paper deals with the training phase (Leibler, 2024).

The paper is guided by three research questions: (1) Is the current copyright law adequate to address all of the extensive and transformative copying involved in AI training? (2) To what extent do the doctrines of fair use in the United States and fair dealing in the European Union and India apply? (3) What practical measures should India adopt to reform its regime in accordance with global standards?

II. Data Mining and Copyright Violation: Legal Perspectives from European Union and United States of America

For the purpose to develop trained models that are capable of executing a broad spectrum of tasks, machine learning necessitates an abundance of data. Despite the fact that data or information is not considered subject matter under the copyright law, a substantial portion of the materials used as datasets for machine learning are protected by copyright law, including texts, images, and videos (Kretschmer et al., 2024). Data scraping is a standard initial step in the task of machine learning, which involves the collection and organization of information from online sources. The technologies of mining, learning, and scraping are frequently perplexed, much like the legal frameworks that govern their regulation (Margoni et al., 2022).

Both texts and images are frequently employed as training data. They become viable subjects for protection under copyright as literary or artistic works when expressed in an original form. In general, copyright law distinguishes between protected and unprotected content. Original expressions in literature and artistic creations are represented by the former. The latter is a broad category that does not warrant protection for a variety of reasons, including the absence of originality, the absence of a stable and objective expression, and the expiration of the protection term (Kretschmer et al., 2024).

The central question in the ongoing global debate, including in the United States and the European Union, is whether the use of copyrighted works to train generative AI models constitutes copyright infringement or is protected under exceptions for transient and temporary copying and text and data mining in the EU, or under the fair use doctrine in the United States. EU legislators have mandated that providers of generative AI systems must publicly disclose a summary of the utilization of training data protected by copyright law. This is notwithstanding the fact that the European Union AI Act is not explicitly designed for copyright law.¹ The processes necessary to develop an AI

¹ Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No. 300/2008, (EU) No. 167/2013, (EU) No. 168/2013, (EU)

system are the primary concern from an input perspective. The scraping and extraction of relevant information from foundational datasets, which frequently contain copyrighted materials, is a necessary step during the training phase of AI systems. In the EU, these activities are primarily governed by two text and data mining (TDM) exceptions outlined in the 2019 Copyright in the Digital Single Market (CDSM) Directive,² which address TDM for scientific purposes (Art. 3) and what is referred to as “commercial” TDM (Art. 4).

Further, the EU AI Act requires that providers of general-purpose AI models disclose summaries of the data used for training in order to assure compliance with existing EU copyright laws and enhance transparency (Guadamuz, 2025). Nevertheless, the AI Act demonstrates deficiencies in this context, including deficient provisions for transformative uses that are comparable to the fair use doctrine and ambiguous data disclosure requirements. Furthermore, the requirement for non-EU AI providers to adhere to EU copyright standards entails substantial compliance burdens, which have the potential to restrict the accessibility of AI technologies within the EU market and stifle innovation. It is imperative to further refine these limitations in order to strike a balance between the facilitation of AI development and the preservation of intellectual property (Leibler, 2024).

TDM is described under the CDSM Directive as any automated analytical method intended to examine digital text and data in order to provide information, such as patterns, trends, and correlations. This extensive definition will potentially include various training activities crucial for the creation of AI systems, especially machine learning types, including generative AI systems, owing to its clarity (Margoni et al., 2022).

2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). European Union. Available at: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng> [Accessed 09.06.2026].

² European Parliament and of the Council. (2019). Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. European Union, <https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng> [Accessed 09.06.2026].

The CDSM Directive stipulates two exceptions related to TDM in Art. 3 and 4. Article 3 lays down an exemption for research organizations and cultural heritage institutions to engage in acts of TDM for the purpose of scientific research, which encompasses both natural and human sciences. This exemption is contingent upon the fact that the organizations must have lawful access to the copyrighted works or subject matter in question, as well as the fulfilment of several additional conditions.

Further, Art. 4 establishes an exception for the reproduction and extraction of lawfully accessed works/subject matter. The objective is to grant legal certainty for actions that may not satisfy the “impermanent and transient copy exception” requirement stipulated in Art. 5(1) of the InfoSoc Directive. In the case of work that is publicly available online, the new exception is subject to reservation by rights holders, including the use of metadata and the terms and conditions of a website or service. Such a reservation shall not impact the application of the TDM exception for scientific purposes under Art. 3 (Dermawan, 2024).

Both TDM exceptions have been criticized for their restrictive nature, which may result in the exclusion of a multitude of applications in this field, particularly those regarding the advancement of AI technologies. However, the commercial TDM exception has regained momentum as a viable, scalable policy option to address generative AI. This is due to the emergence of generative AI and its clash with the copyright domain, as well as a favorable political landscape and timing (Margoni et al., 2022).

As reflected in Germany under Sections 44b and 60d of UrhG (the German Copyright Act), the Art. 3 and 4 of the CDSM Directive, highlight the potential and difficulties of text and data mining exceptions, particularly in relation to AI training. The legal framework is designed to balance the interests of copyright holders with the needs of researchers and commercial entities involved in TDM. Although scientific research is broadly protected under these provisions, commercial AI training remains less well-defined, as it relies on explicit opt-outs and machine-readable notices to determine lawful access. This structure was examined and substantively clarified on 27 September 2024 in

the Hamburg Regional Court's judgment in *Kneschke v. LAION e.V.*,³ the first major court decision on the EU TDM exceptions in the AI context. The *Kneschke* decision established that such protection extends to preparatory work that may not directly generate research findings but constitutes a necessary and essential step in the scientific research process. In that case, the Hamburg Court found that *LAION e.V.*, a non-profit organization, qualified as a research organization authorized to conduct TDM for dataset creation intended for subsequent AI research by others, provided the organization operated non-commercially and commercial partners did not exert decisive influence over research findings (Leistner and Antoine, 2025).

A German approach exemplifies a wider European tendency in which TDM exceptions are intended to promote innovation while preserving intellectual property rights. However, the effective enforcement of these exceptions continues to be challenging, as the legal framework remains in development and the definition of what comprises "lawful access" is the subject of a continuing debate (Rosati, 2025). The problem is considerably more complex due to the necessity of establishing a clear distinction between TDM and AI training. This is because AI training involves additional copyright considerations that are not entirely encompassed by the existing exclusions. In addition, Art. 4(3) of the CDSMD's opt-out mechanism, intended to protect right holders, could unintentionally hinder AI progress by making things legally unclear and hard to work for AI developers (Ziaja, 2024). This shows how important it is to have a more nuanced regulatory approach that works for both non-commercial and commercial purposes. One way to do this may be to employ a hybrid model that combines statutory exceptions for non-commercial usage and opt-out clauses for commercial applications (Thongmeensuk, 2024).

The legal issue in the United States, unlike the EU, is whether such artificial intelligence training operations constitute fair use in light of the lack of a specific TDM exception. Under the judgments given in the

³ Robert Kneschke v. LAION e.V., Case No. 310 O 227/23 (Hamburg Regional Court, Germany, 27 September 2024).

cases *Authors Guild v. HathiTrust*⁴ and *Authors Guild v. Google*⁵ the doctrine of fair use in the United States has been contended to permit a broad spectrum of text and data mining activities involving copyrighted works. The bottom line is that U.S. copyright law may be among the most lenient globally regarding TDM activities, especially when compared to jurisdictions that impose more stringent exceptions and limitations, such as the EU. Furthermore, the Generative AI Copyright Disclosure Bill of 2024 (Leibler, 2024) which was introduced in the House of Representatives and is presently being reviewed by the House Judiciary Committee in the United States proposed that AI developers must send a notice to the U.S. Copyright Office concerning the uses of copyrighted materials in their training datasets. This bill is akin to the EU AI Act in that it prioritizes transparency and disclosure. Its primary objective is to guarantee that copyright holders are informed about the utilization of their works in generative AI systems (Leibler, 2024).

The evolving legal framework concerning fair use, particularly following the decision of Supreme Court of the United States 2023 decision in *Andy Warhol Foundation v. Goldsmith*⁶ and the 2025 decision by United States District Court for the District of Delaware in *Thomson Reuters v. ROSS Intelligence*,⁷ has significant implications for the permissibility of training artificial intelligence systems on copyrighted material. The Court limited the definition of transformative use in the

⁴ *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014). The case is concerned with the issue that whether the act of HathiTrust Digital Library to digitise the copyrighted works for access and full text search for visually impaired users, can be called as fair use or not. The Second Circuit Court affirmed that creating a searchable database and providing access to print-disabled individuals were fair uses under U.S. copyright law. The case mark a significant victory for fair use and library digitization efforts.

⁵ *Authors Guild v. Google*, 804 F.3d 202, 217 (2d Cir. 2015). The Authors Guild and other plaintiffs alleged that practices by Google to digitise books infringed upon their copyrights. The Second Circuit Court affirmed that digitization and display of snippets by Google constituted fair use under 17 U.S.C. § 107. The court emphasized the transformative nature of these uses, noting that they served different purposes from the original works and did not substitute for them. The court acknowledged the commercial nature of Google but also concluded that the public benefits of the project outweighed any potential harm to the rights holders.

⁶ *Andy Warhol Foundation v. Goldsmith*, 598 U.S. 508 (2023).

⁷ *Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc.*, No. 1:20-cv-613-SB (D. Del. 11 Feb. 2025) (Memorandum Opinion).

Warhol case. It ruled that a secondary use does not qualify as transformative solely because it adds a new meaning or aesthetic if the commercial objective is quite similar to that of the original. The Court weakened reliance on broad technological transformativeness as a defense for AI training by rephrasing earlier judgements like *Authors Guild v. Google* to emphasize commerciality and the specific goal of the usage (Lin, 2023).

This narrower view of what is transformative is shown in the *ROSS* judgement. The Delaware District Court determined that reproducing Westlaw headnotes for the purpose of training a competing AI-based research tool was not considered transformative, emphasizing that the use duplicated the expressive content of the headnotes. The court stressed how the usage would hurt Westlaw's current subscription market and other licensing markets that may reasonably be expected to happen. It decided that these factors made the use not fair use. The court also refused to accept comparisons to cases of intermediate copying that involved functional code. It declared that headnotes had creative expression that an AI system may copy or use (Sookman, 2025). When read together, these cases show that U.S. courts are paying more attention to how the market works, what motivates creators, and the specific functional role of copyrighted inputs in AI training. This makes it harder for commercial machine learning systems to use data without permission.

The above approaches also expose gaps and lingering doctrinal issues vis-à-vis the applicability of traditional fair use principles to AI training. In *Warhol* case, the Court limited its analysis to the narrow question of licensing. It left unanswered the question that whether large-scale ingestion of copyrighted works for model training (an upstream activity that does not itself generate output for consumers) should be treated differently from downstream commercial licensing. This raises a fundamental question: does the fair use analysis conflate "use in training" with "use in final outputs," hence imposing an excessively stringent requirement on a phase of AI development that is arguably comparable to intermediate copying? The *ROSS* decision compounds the situation by attributing considerable importance to possible licensing markets, suggesting that copyright holders can create future markets to undermine fair use assertions. If courts persist in endorsing such arguments, the doctrine may evolve into a framework where the

mere potential for a licensing market is suffices to preclude fair use, consequently obligating AI developers to operate within a completely licensed training environment, irrespective of the nature or intent of the use. The dismissal of the intermediate-copying argument also warrants examination, as its rationale may not readily apply to other categories of inputs such as structured datasets, metadata, or embedded content that may operate more as facts or functional systems rather than as creative expressions. Furthermore, the distinction established in ROSS between generative and non-generative AI introduces significant ambiguity, i.e., if generative models produce novel expressions, will courts afford them greater leniency under Warhol, or will the increased emphasis on market harm be applied more rigorously? These unresolved issues underscore that the existing jurisprudence, although safeguarding rightholders, may still lack a consistent and technologically nuanced framework for assessing AI training within the scope of fair use.

In addition of the above, Katharina de la Durantaye in her work has done a comprehensive examination of how various jurisdictions address the use of copyrighted content for training generative AI models (Durantaye, 2025). A tabular representation is provided below to present a global perspective.

Table 1

Juris-diction	Control by Rights Holders	Compensation Mechanisms	Transparency Requirements	Legal Certainty	Notable Features
United States	Fair use doctrine allows certain uses without explicit consent	No statutory compensation; relies on fair use assessments	Minimal; transparency not mandated	Fair use is flexible but can be unpredictable	Fair use provides broad leeway for AI training, but outcomes can vary case by case
Canada	Fair dealing permits specific purposes	No statutory compensation; depends on fair dealing applicability	Not specifically addressed	Fair dealing is more restrictive than U.S. fair use	No specific TDM exception; fair dealing applies to limited purposes like research and education

United Kingdom	Proposed opt-out mechanism for rights holders	Potential for compensation through licensing; details under consultation	Proposals include transparency obligations for AI developers	Legal framework under development	Government proposes allowing AI training on copyrighted works unless creators opt out. (However, no approved policy currently allows AI training unless creators opt out, it was only proposed)
European Union	Rights holders can opt out of TDM exceptions	No direct compensation; opt-out mechanism allows control	Transparency obligations under the AI Act	Clear rules under CDSM Directive and AI Act	Favourable conditions for AI companies; legal certainty provided by specific directives
Israel	Fair use applies	No statutory compensation; relies on fair use assessments	Not specifically addressed	Fair use provides some flexibility (no explicit case law yet)	No specific TDM exception; fair use is somewhat similar to U.S. model
China	Rights holders have significant control	No statutory compensation; unauthorized use is prohibited	Transparency requirements are stringent	Clear but restrictive legal framework	AI training on copyrighted works requires explicit permission; strong enforcement mechanisms
Singapore	Specific exception for computational data analysis	No compensation required under the exception	Transparency not mandated under the exception	Clear provisions under the Copyright Act	Exception allows AI training for both commercial and non-commercial purposes with lawful access
Japan	Broad exception for information analysis	No compensation required under the exception	Transparency not specifically addressed	Clear legal framework supports AI training	Permits use of copyrighted works for AI training without consent, fostering innovation

III. Technical Perspective

From a technical perspective, there are two aspects that require deliberation regarding copyright violations. These are, first, the various stages of training AI models and copyright violations associated with each state and second, the process of converting the training data into synthesised content.

The various stages of AI model training and the associated copyright concerns are outlined below:

Stage	Copyright Issues
Stage 1: Data Collection & Organization	Sourcing often involves copyrighted material (texts, images, etc.), raising reproduction and adaptation rights concerns
Stage2: Training	Involves reproduction of works, potentially infringing unless covered by exceptions or licenses
Stage 3: Deployment & Use	Model outputs may resemble protected content; traditional infringement rules may apply

From the second perspective, AI models have advanced significantly in their ability to generate a diverse range of synthesized content, including text, images, graphics, audio, and code. For example, advanced image-generation models, such as diffusion models (DMs), can produce highly realistic and intricate images and artworks.

Likewise, large language models (LLMs) such as ChatGPT, can be utilised to generate cohesive and imaginative textual content across diverse genres and narratives. These advanced models are termed “Deep Generative Models” (DGMs). Nonetheless, due to the exceptional fidelity and authenticity of the content produced by DGMs, questions have emerged over the related copyright issues (Ren et al., 2024).

In this section, an overview of present computational approaches proposed for copyright protection are discussed from a technical standpoint. These are feasible alternatives for both AI model creators and users regarding copyright protection. Generally, these computational techniques can be classified based on the copyright holder as below mentioned.

Firstly, for the “source data owners.” They can protect their original works by: (a) “crafting unrecognisable examples,” which involves

manipulating data to obstruct DGMs from extracting information; (b) employing “watermark techniques” to trace and identify whether a generated work is derived from their original creation; (c) utilising “machine unlearning” to request the deletion of their data from the model or its output upon detecting copyright infringement; (d) implementing “dataset de-duplication” to eliminate redundant data, thereby reducing the memorisation effect and preventing the generation of training data; (e) applying “alignment,” which employs a reward signal to diminish memorisation in LLMs; and (f) exploring other methods, including enhanced training and generation algorithms for improved LLM behaviour (Ren et al., 2024).

Secondly, for users who create new works with the help of DGMs (Deep Generative Models), the protected work constitutes the content generated by these models. Copyright for such content does not depend much on how the DGM created it, since traditional copyright methods can still be used to protect it. To identify and trace ownership, DGM providers often use “watermarking strategies” by embedding invisible marks in the content or the model itself (Ren et al., 2024).

Illustratively, consider the programs like Stable Diffusion, Midjourney and DALL-E 2 that use generative AI to create new visuals. In each program, the end user submits a textual description (i.e., prompt) specifying the desired image, and via a complex process, the program produces a new image that aims to correspond with the given description. The AI picture generating program relies on a ‘text encoder’ and an ‘image decoder.’ The text encoder converts the prompt into a ‘transition output.’ The latent output can subsequently be decoded by the image decoder into images. Neither the image encoder, the latent representation, nor the image decoder contains images in themselves (Shoemaker, 2024).

It is also pertinent to discuss the lawsuit of GitHub Copilot. GitHub Copilot was developed by GitHub in collaboration with OpenAI. An AI-powered code completion tool assists developers by suggesting code snippets as they type. It leverages upon the Codex model of OpenAI that is trained on a big corpus of publicly available code, including the repositories hosted on GitHub. While Copilot aimed to enhance productivity, it was surrounded in legal controversies concerning copyright

and open-source license compliance. In November 2022, a class-action lawsuit was filed against GitHub, Open AI and Microsoft, alleging that GitHub Copilot infringed copyright laws by reproducing code from open-source projects without proper attribution. The lawsuit claimed violations of the Digital Millennium Copyright Act, open-source licenses, and other legal frameworks. In July 2024, a California court allowed only two claims to proceed pertaining to violation of open source license and breach of contract. The court found that the code generated by Copilot was not sufficiently similar to the original works so as to constitute copyright infringement.

The most appropriate precedent on the above situation is the Authors Guild v. Google case.⁸ Herein, as discussed above, the Second Circuit Court of Appeals held that Google's act of scanning millions of copyrighted books for computer analysis and indexing their contents to provide text snippets in response to user search requests constituted fair use. The court also held that the use of copyrighted books was transformative, since the corpus further enhanced access to information. Google replicated entire books to fulfil its revolutionary objective of indexing their contents for computational analysis and search. Moreover, since Google supplied only three concise excerpts from each book, the snippets were unlikely to adversely affect the market for the books.

IV. Legal Challenges in Training AI: From Data Mining to Copyright Infringement

Several fundamental questions give rise to the legal disputes about data use in machine learning. Firstly, there is significant ambiguity regarding the legal status of input data, which complicates the use of data without risking copyright violations. Many datasets for training models contain copyrighted content that needs permission under copyright law, which creates challenges for creators. Furthermore, the exemptions permitted for text and data mining or research might not cover every company operation, leading to uncertainty about legal rights. By making access dependent on right-holder approval, the EU's legal ac-

⁸ Authors Guild v. Google, 804 F.3d 202 (2nd Cir. 2015).

cess criteria might ironically stifle innovation and create even more uncertainty for machine learning applications. Finally, obligations for artificial intelligence companies to reveal their training data could add more administrative load and complexity, stressing the need for more precise legal frameworks to promote innovation and copyright protection (Kretschmer et al., 2024).

The crux of the discussion is the legality of data mining from copyrighted material and whether this practice qualifies as fair use or violates intellectual property rights. The implications of data mining from copyrighted material on the concept of fair use in intellectual property law are significant and multifaceted. The ongoing debate, exemplified by the Getty Images and Stability AI case, highlights the tension between the need for large datasets in AI training and the rights of copyright holders, raising questions about whether such uses can be classified as fair use or infringement (Lintang Kanaya and Santoso, 2025). In the U.S., the fair use doctrine emphasizes transformative use, yet recent analyses suggest a shift towards considering market harm, which could complicate the legal landscape for generative AI outputs. Conversely, the European Union has introduced specific exceptions for text and data mining, particularly for scientific research, reflecting a more permissive approach to balancing copyright protection with innovation needs (Fernández-Molina and De La Rosa, 2024). This divergence in legal frameworks underscores the necessity for adaptive copyright laws that accommodate the evolving digital environment and the growing importance of data mining in different fields.

A rigorous analysis indicates that neither the TDM exceptions in the EU nor the U.S. fair use doctrine consistently address the conflict between innovation and market detriment. The *Getty Images v. Stability AI*⁹ case in the UK shows that arguments about copyright infringement must make a clear difference between direct reproduction and the wider, secondary effects of AI model output. In the U.S., on the other hand, the *Concord* case¹⁰ shows how contributory and vicarious liability can apply to the systematic use of protected works. It is clear that there

⁹ *Getty Images (US) Inc & Ors v. Stability AI Limited*, EWHC 2863 (Ch) (2025).

¹⁰ *Concord Music Group v. Anthropic*, — Cv-01092 (M.D. Tenn. 18 Oct. 2023) (2023).

needs to be a “intermediate copyright exception” just for AI training. This exception should require clear transformation, a market substitution analysis, and transparency.

IV.1. Whether Infringement or not? A Jurisprudential Perspective

Although copyright law delineates explicit regulations about copying, reproduction, and fair use, these principles were established before to the advent of AI and may not adequately fulfil the justifications of copyright law and philosophy in their application to AI-related use cases. The application of classical doctrines to AI learning necessitates a re-evaluation of essential copyright concepts, especially concerning the transformative use of data in AI models, considerations of market harm, and the overarching objectives of copyright law.

The reproduction right, conferred by U.S. law, is among the most fundamental aspects of copyright. AI training frequently entails the reproduction of works — be it photos, text, music, or other copyrighted materials — enabling the analysis of these works for patterns by machine learning algorithms. Initially, this reproduction may appear to constitute copyright infringement due to the unauthorized duplication involved. The characteristics of reproduction in AI training contest conventional concepts of infringement. AI models do not utilize these works for their emotive content, as a person would; rather, they employ them for the underlying data. Machine learning algorithms analyze works by breaking them down into data points, patterns, and relationships, which are subsequently utilized to train the model. This prompts a vital inquiry: does this form of usage qualify as copying as defined by copyright law (Price, 2022)?

Copyright law safeguards the expression of concepts, not the concepts themselves. AI training often emphasizes the conceptual or informational elements of works, prioritizing data extraction over the reproduction of expressive content. This contrast corresponds with overarching copyright theories, especially utilitarianism, which posits that copyright serves to reconcile the protection of authors with the advancement of the collective benefit through the spread of innovative information (Lemley, 2015).

A significant defense for AI engineers is the idea of fair use, especially its focus on transformative use. As described earlier, in instances such as *Authors Guild v. Google*, the court determined that Google's digitization of books to establish a searchable database was significantly transformative, as the application diverged from the original intent of the works. Likewise, AI training may be regarded as transformative since it employs works for data extraction instead of mere expressive replication. Nonetheless, not all AI training can be considered transformative.

The degree to which AI training is deemed transformational hinges on whether the model merely replicates the original work or employs it for a markedly distinct objective. In the case of *Concord Music Group v. Anthropic*, the plaintiffs assert that Anthropic's Claude model was trained on copyrighted songs. If the model were explicitly built to replicate these lyrics or provide outputs that closely mimic them, the case for transformative use appears to diminish. Conversely, generalized training, when the AI is exposed to a diverse array of resources to cultivate broad competencies, may provide a more compelling argument for transformational usage.

From this viewpoint, generalized training can be likened to human learning, wherein exposure to diverse materials underpins the emergence of novel creativity. Ecclesiastes' declaration "there is nothing new under the sun" and Newton's characterization of "standing on the shoulders of giants" profoundly resonate in this context. The overarching objective bolsters a transformative use claim; yet the unpredictability of fair use doctrine implementation may introduce ambiguity that could impede innovation.

IV.2. Philosophical Approaches to Copyright and AI Training

In addition to legal principles and litmus tests, diverse views of copyright law provide unique viewpoints on the legitimacy of AI training. The natural rights theory, as proposed by John Locke, posits that creators possess intrinsic rights to govern and derive benefits from their works. This viewpoint suggests that use copyrighted resources for AI training without specific consent infringes upon these essential rights,

irrespective of the intent or result (Locke, 1823). Conversely, utilitarian theories, shaped by philosophers such as Bentham and Mill, underscore the broader societal advantages of permitting AI systems to utilize huge datasets, even protected materials. Advocates contend that such access stimulates innovation, augments technological competencies, and tackles significant societal issues, like bias reduction and the enhancement of decision-making processes. This approach promotes a balance between safeguarding creators' rights and advancing the public good through technological innovation. Theory of personality, linked to Hegel, asserts that creative works are manifestations of the creator's identity. AI training that replicates an individual's distinctive creative style or voice may violate the personal rights of creators, as evidenced by lawsuits such as *Concord Music Group v. Anthropic*. This approach emphasizes the significance of honoring the personal relationship between creators and their works, even in cases of transformative replication (Leibler, 2024).

Beyond Locke, Bentham, and Hegel, a relational perspective should guide the formulation of copyright policy for artificial intelligence. Rawlsian principles of distributive justice (Freeman, 2018) necessitate regulatory frameworks that fairly reconcile collective advancement with creator autonomy, particularly in contexts where AI broadens access to knowledge while potentially undermining individual expression (Gabriel, 2022; Liu, 2012). Applying digital constitutionalism and commons theory (De Gregorio, 2022), influenced by thinkers like Lawrence Lessig (1999; 2001) and James Boyle (2018), further reinforces the conceptualization of data as a semi-common (shared resource), thereby justifying graduated exceptions for transformative AI training in conjunction with comprehensive remuneration for creators (Warso and Tarkowski, 2024).

IV.3. Critical Legal Perspective

The existing copyright law appears inadequate in addressing the intricacies of AI training, especially the differentiation between generalized and tailored training. Copyright law fundamentally aims to reconcile the rights of authors with the overarching public interest in the

distribution of knowledge and ideas. Historically, copyright law has acknowledged the significance of enabling individuals to learn from and expand upon preexisting works. The idea-expression distinction demonstrates that copyright safeguards the particular expression of ideas, excluding the concepts from protection. Nonetheless, depending exclusively on fair usage for AI training has certain issues.

The absence of legal certainty might deter innovation, especially for smaller developers who may not possess the means to manage the intricacies of copyright litigation (Leibler, 2024).

AI training, especially when aimed at identifying patterns and data points within extensive datasets, frequently concentrates on these unprotectable components rather than the expressive substance of creations. From this viewpoint, AI training can be regarded as an integral aspect of the inherent processes of learning and invention, endeavors that copyright law has historically aimed to encourage (Lemley, 2015). Moreover, a licensing-based method, although ostensibly simple, prompts apprehensions around possible restrictions on training data. Dependence on licensing may confine training data to public domain materials, a corpus that may not accurately reflect modern society and potentially reinforce biases in AI systems. This underscores the necessity to investigate alternate legal frameworks that guarantee access to varied training data while safeguarding authors' rights (Gravett, 2020).

Legal uncertainty on “temporary copies” e.g., *Infopaq International A/S v. Danske Dagblades Forening* (2009)¹¹ directly affects AI training. In *Infopaq International A/S v. Danske Dagblades Forening*, the Court held that even very short fragments reproduced transiently in the course of a technical process may constitute a “reproduction in whole or in part.” This ruling created a restrictive baseline for EU law, meaning that temporary copies generated during automated processes — including TDM and AI training — may require authorization unless an exception applies. The Supreme Court of India in *R.G. Anand* reaffirmed the idea — expression dichotomy. It held that infringement occurs only when the “substantial and material” expression is copied,

¹¹ *Infopaq International A/S v. Danske Dagblades Forening*, (C-5/08) EU:C:2009:465 (2009).

and the ordinary observer would unmistakably identify the later work as a copy. Copying of ideas, themes, or general concepts is not infringement. This is a foundational principle in Indian copyright law. This insight supports the legitimacy of transformative, data-driven learning where outputs are not substitutive of the original. Policymakers should accordingly consider a graduated access model with tiered dataset permissions based on the degree of transformation and commercial impact.

IV.4. Suggestions: Technological methods and Strategies

The following technological strategies help protect original content from misuse by DGMs through prevention, tracing, removal, and responsible model training practices.

IV.4.1. Unrecognisable Examples

The objective of this technique is to inhibit models from acquiring notable characteristics of the source works. This strategy entails altering the dataset so that the training samples lack identifiable characteristics. For example, data proprietors can implement tactics such as introducing noise, blurring, or modifying colours to hinder DGMs from extracting valuable information. The result frequently yields inferior or entirely dissimilar outputs from those intended, thereby safeguarding the original content from direct reproduction (Shan et al., 2025).

IV.4.2. Watermarking Techniques

The purpose of watermarking technique is to monitor the utilisation of original content and enable owners to substantiate ownership. Watermarks may be embedded within images or various data formats in a manner that is undetectable to human observers yet discernible by specialised algorithms. These watermarks assist data proprietors in verifying whether their data has been utilised for training AI models, allowing them to address any unauthorised usage. It serves as a deterrent to unauthorised duplication (Christ et al., 2023).

IV.4.3. Machine Unlearning

The primary objective of machine unlearning is to enable data proprietors to “remove” their contributions from a model. This entails a process aimed at inducing a trained model to forget a certain subset of its training data. Upon a data owner’s request for content removal, the model parameters are modified to simulate a condition where that item was never part of the training dataset. This is especially crucial for maintaining continuous protection, even subsequent to the training of models with the data (Chen and Yang, 2023).

IV.4.4. Dataset De-duplication

The aim is to decrease the likelihood of memorization by DGMs, thereby diminishing the chance of copyright violation. Data proprietors may establish procedures to eliminate redundant entries from the training datasets prior to their utilization. This ensures the model does not acquire redundant features. Reducing redundancy diminishes the likelihood that the model will unintentionally produce outputs that closely resemble the original works (Webster et al., 2023).

IV.4.5. Alignment Strategies

It aims to inhibit rote memorisation during model training. Employing reinforcement learning methodologies in which the model receives feedback to emphasise original traits while disregarding memorised specifics constitutes a method of alignment. This technique promotes the creation of generative content that is inspired by, rather than directly replicated from, the training data.

IV.4.6. Improved Training Algorithms

It endeavours to promote lawful model training that adheres to copyright regulations. Researchers are always advancing novel algorithms and approaches that emphasise responsible model training.

The paper presents a novel contribution by illustrating how conventional copyright theories and general legal principles can be me-

thodically integrated with technical interventions that influence them. By synthesising these elements, it establishes a coherent theoretical framework that has been largely lacking in existing scholarship. The following framework contextualises the same.

Table 2

Theory of Copyright Law	Core Principle	Implications for AI Training	Relevant Technical Interventions	Justification / Explanation
Natural Rights Theory	Creators have inherent rights over their creations	Unauthorized use of works violates moral claims of authors	Machine Unlearning, Watermarking Techniques	Unlearning allows for revocation of consent post-training, aligning with authorial control. Watermarking ensures moral ownership is traceable
Utilitarian Theory	Promote greatest public good through knowledge dissemination	Limited use may be justified if it leads to social benefit	Alignment Strategies, Improved Training Algorithms	These promote ethical, generalized learning that reduces harm while enhancing innovation — a utilitarian objective
Personality Theory	Works reflect the author’s identity and spirit	Replication of style or voice harms creator’s personality rights	Unrecognizable Examples, Dataset Deduplication	De-personalization of training data and filtering duplicates limit replication of stylistic signatures
Economic/ Market-Based View	Protects commercial interest and market viability of works	Uses that substitute or devalue original works harm markets	Dataset Deduplication, Watermarking Techniques	Reduces direct substitution risks; watermarking assists in enforcement of market-related claims

Fair Use Doctrine (U.S.A.)	Allows limited, transformative use if it serves public interest without harming rights	Certain AI training qualifies as fair use based on purpose, effect, and transformation	Alignment Strategies, Improved Training Algorithms	These prioritize transformation and abstraction, aiding fair use arguments by ensuring outputs are not derivative
Idea-Expression Dichotomy	Protection applies to expression, not facts or concepts	AI typically extracts data points, not expressive content	Improved Training Algorithms, Dataset Deduplication	Technical focus on data generalization and abstraction aligns with non-infringing idea-level usage
Transformative Use (subset of Fair Use)	New use with different purpose or meaning may be permitted	Generalised training is more likely to qualify than replication	Alignment Strategies, Unrecognizable Examples	Promotes transformation over reproduction; suppresses surface-level similarities
Access to Knowledge and Learning	Balances author incentive with public benefit in learning	AI as a learning mechanism should be supported legally	Improved Training Algorithms, Machine Unlearning	Promotes responsible access while ensuring retractability of content in sensitive cases

IV.5. Proposed Changes in Legal Framework

In addition to the aforementioned technological strategies, several legal proposals are advanced by the author to address the complexities arising at the intersection of AI and copyright.

First, the proposed “Fair Learning” doctrine represents an extension of the existing fair use framework, specifically tailored to machine learning (ML) systems. Conventional fair use frequently encounters challenges in accommodating the distinctive data utilisation patterns

in AI, where context, predictability, and intent diverge from traditional copyright scenarios (Gravett, 2020).

Second, the proposed “Presumption of Fairness” approach would grant AI developers a rebuttable presumption of fairness when using data elements that are not protected by copyright, including ideas, facts, and functional aspects of works. This assumption recognises that although the AI may utilise copyrighted materials for learning, it does not constitute infringement unless it replicates the protected speech (Gravett, 2020).

Third, the proposed “Data Mining Safe Harbor Provisions” would establish legal protections shielding developers from liability in specified circumstances, thereby encouraging data use without fear of infringement. This means defining unambiguous boundaries inside which developers can operate and train their models. According to the proposed safe harbor, an AI system may not infringe if it uses portions of a literary work only for analysis or idea extraction. Such rules will reduce legal uncertainty for smaller businesses or start-ups, hence allowing them to engage with large databases free of litigation risk. It might be organised to ensure that although developers would not need permits for every data used in training, they must nonetheless respect the rights of authors by avoiding direct copying of their original content for business reasons (Gravett, 2020).

Fourth, “Statutory AI Training Exceptions” may be considered. Although some jurisdictions address AI training under existing fair use doctrines or text and data mining exceptions, there is a growing call for narrowly tailored, clearly defined statutory exceptions that balance innovation with the protection of creators’ rights. Implementing opt-out mechanisms, such as metadata tagging or licensing tools like Creative Commons AI licenses, empowers rights holders to control the use of their works (Buick, 2025).

Finally, as a matter of policy reform, the establishment of centralized licensing platforms may be considered. Such platforms would manage and license copyrighted works, monitor usage, ensure compliance, and distribute royalties to rights holders on the basis of actual use. By implementing revenue sharing models similar to music rights management, these platforms can provide fair compensation to creators.

IV.6. Proposed Integrated Framework

The study suggests an Integrated Framework for AI Training and Copyright Governance. The approach is intended to connect doctrinal principles, legislative instruments, and regulatory outcomes into a cohesive model. At the centre of this framework lies the Fair Learning Doctrine, based on a systematic four-factor assessment (i) purpose of use, (ii) nature of source material, (iii) extent and significance of extraction, and (iv) potential market impact. It is designed to distinguish socially advantageous machine learning practices from exploitative or substitutive copying.

The framework will also integrate safe-harbour provisions to ensure transparent, accountable, and non-substitutive AI training procedures, along with opt-out mechanisms that uphold the autonomy of right holders in commercially sensitive situations. This hybrid approach regards AI training datasets as semi-commons infrastructure, harmonizing the interests of creators with the public benefit derived from data-driven innovation.

IV.7. Lessons for India

Under the present regime, Section 52 of the Indian Copyright Act (1957) establishes a limited enumeration of exceptions to copyright infringement, including fair dealing for purposes of research, private study, criticism, and news reporting. The scenario stems from Anglo-Saxon copyright traditions, which provides predictability but constrains adaptation to emerging technological challenges. The doctrinal gap is evident: Section 52 contains no express provision authorizing text and data mining or artificial intelligence training activities. The purpose of Section 52 includes research, criticism, news reporting, and similar traditional uses. They were drafted in an era preceding generative AI and do not contemplate the characteristics of machine learning systems. While it can be argued that “research” might be interpreted broadly to encompass AI training, such interpretations remain contested and uncertain, leaving developers, rights holders, and courts without clear statutory guidance. This uncertainty is currently being tested in ANI

Media Pvt. Ltd. v. OpenAI, pending before the Delhi High Court, in which ANI alleges that OpenAI used its copyrighted news content without authorization to train ChatGPT (Jha and Bernd, 2025). OpenAI, in response, has raised jurisdictional challenges arguing that its data storage and training occurs outside India. Meanwhile, *amicus curiae* have called for a two-step assessment: 1) determining whether the use falls within a Section 52 exception, and 2) conducting a fairness analysis (Jha and Bernd, 2025).

Accordingly, to address the doctrinal gap in Section 52 of the Indian Copyright Act (1957) the introduction of a “Fair Learning Exception” is proposed. This exception would expressly authorize AI training activities, subject to carefully calibrated and clearly defined conditions. Unlike traditional fair dealing, this exception would recognize AI training as a distinct form of use, permitting copying for transformative computational analysis while maintaining robust copyright protection for works outside the scope of exception.

The exception should be guided by core principles balancing innovation with creator rights. AI training must serve genuinely transformative purposes (i.e., extracting patterns or knowledge rather than reproducing creative expression) and it should not substitute the original works. The Fair Learning Doctrine can operationalize this through a four-factor test examining the purpose and character of training, the nature of copyrighted works, the quantity used, and the impact on market value.

Rights holder’s interests should be protected through a centralized opt-out registry, enabling them to exclude works from AI training datasets. Developers must search the registry before commencing training and exclude registered content. Transparency and accountability obligations would require detailed documentation of dataset composition, curation methods, technical safeguards against unauthorized memorization, and compliance with the registry, accessible to the Copyright Board and affected rights holders.

Commercial deployment must include safeguards preventing outputs that replicate copyrighted works. Developers should monitor

outputs for unauthorized reproduction, remediate any incidents, and maintain accountability. Compliance with international obligations, particularly the three-step test under Berne Convention, should be explicitly incorporated.

The Copyright Board should be empowered to issue interpretive guidelines, manage registry operations, resolve disputes, and oversee enforcement. Technological safeguards like watermarking, dataset deduplication, and machine unlearning should be encouraged. The exception would operate alongside licensing mechanisms, allowing revenue-sharing models while providing a baseline of permissible unlicensed use.

A phased implementation is recommended, beginning with pilot programs designed to build institutional capacity and refine procedural frameworks. The amendment should align with constitutional commitment to freedom of expression and promotion of scientific progress, integrate with the national AI strategy, and reflect international developments while tailoring governance to Indian legal and developmental context. Regular review by the Copyright Board would ensure the exceptions remain effective and adapt to technological and market evolution.

Translating this framework into Indian law necessitates a phased and carefully structured approach. First, targeted amendments to the Copyright Act should clarify the legal status of temporary copies, text and data mining, and non-substitutive AI training. Second, the Department for Promotion of Industry and Internal Trade (DPIIT) should issue regulatory guidelines defining permissible training practices, transparency duties, and metadata standards for dataset documentation.

Finally, the paper proposes a tripartite oversight mechanism comprising: (i) the Copyright Board, responsible for statutory interpretation and the administration of compulsory licensing; (ii) an independent AI Ethics Committee tasked with evaluating technical compliance and transparency standards; and (iii) rights-management platforms to administer remuneration schemes, opt-out mechanisms, and dataset permissions. This institutional structure would anchor the Fair Learning Doctrine within India's broader copyright ecosystem, providing clarity for developers while safeguarding creators' interests.

V. Conclusion

In conclusion, AI training and copyright law represent a paradigm shift in the digital age, necessitating legal principles to adapt to technological advancement.

This paper has examined the complex intersection of copyright law and artificial intelligence training, identifying a critical regulatory gap where existing fair use and fair dealing doctrines prove inadequate. Through analysis of the European Union and United States approaches, this paper argues that neither statutory text-and-data-mining exceptions nor flexible fair use doctrine consistently resolve tensions between AI innovation and creator protection.

The solutions proposed herein combine AI training with copyright to build technologies and protect intellectual property. The fair learning doctrine, safe data mining harbours, and ethical requirements together form an effective framework for the same. The Integrated Framework proposed herein exemplifies that doctrinal principles, technological realities, and policy design can be harmonised to shape a coherent approach to AI training. By adopting them Indian copyright law can evolve into a balanced model that supports innovation without compromising the integrity of creative labour.

References

Boyle, J., (2018). *Public Domain: Enclosing the Commons of the Mind*. Yale University Press.

Buick, A., (2025). Copyright and AI training data — Transparency to the rescue? *Journal of Intellectual Property Law & Practice*, 20(3), pp. 182–192, doi: 10.1093/jiplp/jpae102.

Chen, J. and Yang, D., (2023). Unlearn What You Want to Forget: Efficient Unlearning for LLMs. *arXiv:2310.20150*, doi: 10.48550/arXiv.2310.20150.

Christ, M., Gunn, S. and Zamir, O., (2023). Undetectable Watermarks for Language Models. *arXiv:2306.09194*, doi: 10.48550/arXiv.2306.09194.

De Gregorio, G., (2022). *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society*. Cambridge University Press. DOI: 10.1017/9781009071215.

Dermawan, A., (2024). Text and data mining exceptions in the development of generative AI models: What the EU member states could learn from the Japanese “nonenjoyment” purposes? *The Journal of World Intellectual Property*, 27(1), pp. 44–68, doi: 10.1111/jwip.12285.

Durantaye, K., de la, (2025). Control and Compensation. A Comparative Analysis of Copyright Exceptions for Training Generative AI. *IIC – International Review of Intellectual Property and Competition Law*, 56(4), pp. 737–770, doi: 10.1007/s40319-025-01569-6.

Fernández-Molina, J.-C. and De La Rosa, F.E., (2024). Copyright and Text and Data Mining: Is the Current Legislation Sufficient and Adequate? *Portal: Libraries and the Academy*, 24(3), pp. 653–672, doi: 10.1353/pla.2024.a931775.

Freeman, S., (2018). Rawls on Distributive Justice and the Difference Principle. In: Olsaretti, S., (ed.), (2018). *The Oxford Handbook of Distributive Justice*. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199645121.013.2.

Gabriel, I., (2022). Toward a Theory of Justice for Artificial Intelligence. *Daedalus*, 151(2), pp. 218–231, doi: 10.1162/daed_a_01911.

Gravett, W., (2020). The Dark Side of Artificial Intelligence: Challenges for the Legal System. *Southern African Public Law*, 35(1), doi: 10.25159/2522-6800/6979.

Guadamuz, A., (2025). The EU’s Artificial Intelligence Act and copyright. *The Journal of World Intellectual Property*, 28(1), pp. 213–219, doi: 10.1111/jwip.12330.

Jha, P. and Bernd, J.J., (2025). Does Human Learning equal Machine Learning? High Court of Delhi to rule on lawfulness of TDM for Machine Learning. *Kluwer Copyright Blog*. Available at: <https://legalblogs.wolterskluwer.com/copyright-blog/does-human-learning-equal-machine-learning-high-court-of-delhi-to-rule-on-lawfulness-of-tdm-for-machine-learning/>.

Kretschmer, M., Margoni, T. and Oruç, P., (2024). Copyright Law and the Lifecycle of Machine Learning Models. *IIC – International*

Review of Intellectual Property and Competition Law, 55(1), pp. 110–138, doi: 10.1007/s40319-023-01419-3.

Leibler, Y., (2024). From Infringement to Innovation: Reimagining Copyright for AI Training Datasets (SSRN Scholarly Paper No. 4986763). *Social Science Research Network*, doi: 10.2139/ssrn.4986763.

Leistner, M. and Antoine, L., (2025). TDM and AI Training in the European Union — From “LAION” to Possible Ways Ahead? *GRUR International*, 74(11), pp. 1027–1044, doi: 10.1093/grurint/ikaf114.

Lemley, M.A., (2015). Faith-Based Intellectual Property. *UCLA Law Review*, 62, pp. 1328–1344.

Lessig, L., (1999). *Code and other laws of cyberspace (Nachdr.)*. N.p.: Turtleback.

Lessig, L., (2001). *The Future of Ideas: The Fate of the Commons in a Connected World*. N.p.: Random House.

Lin, P.K., (2023). Retrofitting Fair Use: Art & Generative AI after Warhol (SSRN Scholarly Paper No. 4566945). *Social Science Research Network*, doi: 10.2139/ssrn.4566945.

Lintang Kanaya, A. and Santoso, B., (2025). Analysis of Copyright Implementation in Getty Images and Stability AI as a Case Study of Generative AI. *International Journal of Social Science and Human Research*, 8, doi: 10.47191/ijsshr/v8-i9-49.

Liu, D., (2012). Copyright and the pursuit of justice: A Rawlsian analysis. *Legal Studies*, 32(4), pp. 600–622, doi: 10.1111/j.1748-121X.2012.00235.x.

Locke, J., (1823). *Two Treatises of Government*. In: *The Works of John Locke. A New Edition*. In Ten Volumes. Vol. V. London: Printed for Thomas Tegg; W. Sharpe and Son. Available at: <https://www.yorku.ca/comminel/courses/3025pdf/Locke.pdf>.

Margoni, T., Pedro Quintais, J. and Schwemer, S.F., (2022). Algorithmic propagation: Do property rights in data increase bias in content moderation? Part I. *Kluwer Copyright Blog*. Available at: <https://legalblogs.wolterskluwer.com/copyright-blog/algorithmic-propagation-do-property-rights-in-data-increase-bias-in-content-moderation-part-i/>.

Price, H., (2022). Periodic Leave: An Analysis of Menstrual Leave as a Legal Workplace Benefit. *Oklahoma Law Review*, 74(2), p. 187.

Quang, J., (2021). Does Training AI Violate Copyright Law? *Berkeley Technology Law Journal*, 36(4), doi: 10.15779/Z38XW47X3K.

Ren, J., Xu, H., He, P., Cui, Y., Zeng, S., Zhang, J., Wen, H., Ding, J., Huang, P., Lyu, L., Liu, H., Chang, Y. and Tang, J., (2024). Copyright Protection in Generative AI: A Technical Perspective. *arXiv:2402.02333*, doi: 10.48550/arXiv.2402.02333.

Rosati, E., (2025). Copyright Exceptions and Fair Use Defences for AI Training Done for “Research” and “Learning,” or the Inescapable Licensing Horizon. *European Journal of Risk Regulation*, 16(3), pp. 961–984, doi: 10.1017/err.2025.10035.

Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R. and Zhao, B.Y., (2025). Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models. *arXiv:2302.04222*, doi: 10.48550/arXiv.2302.04222.

Shoemaker, E., (2024). Is AI Art Theft? The Moral Foundations of Copyright Law in the Context of AI Image Generation. *Philosophy & Technology*, 37(3), p. 114, doi: 10.1007/s13347-024-00797-x.

Sookman, M.T.L.-B.B., (2025). AI Training Copyright Infringement and Fair Use: Thomson Reuters v. Ross. *Lexology*. Available at: <https://www.lexology.com/library/detail.aspx?g=b200f794-16f5-4c21-9542-d995517ab87a>.

Thongmeensuk, S., (2024). Rethinking copyright exceptions in the era of generative AI: Balancing innovation and intellectual property protection. *The Journal of World Intellectual Property*, 27(2), pp. 278–295, doi: 10.1111/jwip.12301.

Warso, Z. and Tarkowski, A., (2024). Commons-based Data Set Governance for AI. *Open Future*. Available at: <https://openfuture.eu/publication/commons-based-data-set-governance-for-ai> [Accessed 13.06.2026].

Webster, R., Rabin, J., Simon, L. and Jurie, F., (2023). On the Duplication of LAION-2B. *arXiv: 2303.12733*, doi: 10.48550/arXiv.2303.12733.

Ziaja, G.M., (2024). The text and data mining opt-out in Art. 4(3) CDSMD: Adequate veto right for rightholders or a suffocating blanket for European artificial intelligence innovations? *Journal of Intellectual Property Law & Practice*, 19(5), pp. 453–459, doi: 10.1093/jiplp/jpae025.

Information about the Author

Siddharth Badkul, PhD, Research Scholar at Maharashtra National Law University, Nagpur; Assistant Professor of Law, Bennett University, Greater Noida, India

siddharthbadkul@gmail.com

ORCID: 0009-0001-2297-0409

Received 14.10.2025

Revised 01.02.2026

Accepted 09.02.2026