Christ University Law Journal
2025, Vol. 14, No. 2, *1-21*
ISSN 2278-4322/https://doi.org/10.12728/culj.27.1

Fair-Use Doctrine: Copyright Challenges Posed by AI Generative Technology and In-Text and Data Mining Training

Shridul Gupta*

Abstract

Artificial intelligence (AI) innovators have invested billions in research and development to create advanced software and hardware tools. AI has generated new businesses and start-ups, providing employment to millions. However, despite its transformative potential, AI innovation has not received sufficient support from legal systems and remains constrained by the current intellectual property regime. Fair-use exemptions, particularly with respect to copyright law, have posed challenges for AI development and training. This disconnect arises because copyright law has not evolved in tandem with AI technologies and still reflects principles from an earlier computing era. Consequently, there is a pressing need to examine which provisions of existing copyright frameworks may be impeding AI progress, especially those related to fair-use exceptions. The ambiguity surrounding these exceptions has led to unpredictable judicial interpretations, particularly in the context of AI tools and technologies. As numerous generative AI systems, including OpenAI's ChatGPT, rely on large datasets that incorporate both copyrighted and non-copyrighted materials, the process of AI training has become a focal point of legal, ethical, and artistic debate. This paper explores these complexities and examines the emerging copyright challenges associated with the training and use of generative AI systems.

* Maharaja Agrasen Institute of Management Studies, Delhi, India; shridulgupta82@gmail.com

Keywords: Copyright Act 1957, Machine Learning Systems, Original Work, Training Datasets, Transformative-use Doctrine

1. Introduction

Copyright law traditionally extends protection only to original works of human authorship. Most jurisdictions have implicitly assumed that the 'author' must be a human being. The United States Copyright Office, for example, routinely denies registration to works created without human involvement. In Thaler v. Perlmutter, the U.S. District Court for the District of Columbia upheld the Office's refusal to register an AI-generated image, affirming that 'human authorship is an essential requirement for copyright protection.'[1] In a similar matter, a registration was sought in 2023 by artist Jason Allen for his artwork titled 'Théâtre D'opéra Spatial.' The piece was generated using the Midjourney platform. The Review Board of the U.S. Copyright Office held that the work lacked the requisite human authorship and therefore could not be protected.[2]

These decisions underscore an emerging tension between technological creativity and legal orthodoxy. As AI systems such as ChatGPT and DALL-E increasingly generate expressive works, courts and policymakers must determine whether existing copyright doctrines can accommodate non-human creators. The research problem that this paper examines is whether doctrines like fair use and fair dealing developed for a human-authorship paradigm remain adequate for machine learning and generative-AI training.

2. Doctrine Of Fair Use

The U.S. Copyright Act of 1976 codifies the fair-use doctrine in section 107.[3] It allows certain unauthorised uses of copyrighted material for purposes such as criticism, comment, news reporting, teaching, scholarship, or research. The statute enumerates four non-exclusive factors: (1) the purpose and character of the use (2) the nature of the copyrighted work (3) the amount and substantiality of the portion used and (4) the effect of use upon the potential market.[4]

Fair use functions as a safety valve that reconciles the exclusive rights of authors with the constitutional objective of 'promoting the progress of science

---

1 Thaler v. Perlmutter, No. 1:22-cv-01564, 2023 WL 5333236 (D.D.C. Aug. 18, 2023).

2 U.S. Copyright Office Review Bd., *Second Request for Reconsideration for Refusal to Register Théâtre D'opéra Spatial* (Feb. 21, 2023).

3 17 U.S.C. § 107 (2018).

4 *Id*.

and useful arts.'[5] Its inherently flexible character has enabled courts to adapt it to technological innovation, from photocopying to digital sampling and search-engine indexing. Yet this flexibility also breeds uncertainty. In the context of AI, developers claim that copying large datasets to train machine-learning models constitutes transformative use, while rights-holders view the same conduct as massive infringement.

On the other hand, in India, section 52 of the Copyright Act 1957, provides for 'fair dealing' exceptions, covering private use, research, criticism, and review.[6] Unlike the open-ended U.S. doctrine, India's provision enumerates specific permissible acts. Courts here have occasionally borrowed the U.S. four-factor analysis,[7] but the statutory framework remains more restrictive. This divergence between fair use and fair dealing is pivotal in assessing how different legal systems approach AI-training datasets.

3. Important Judicial Decisions On Fair Use Doctrine

The modern debate over AI and fair use echoes earlier disputes over new technologies. A central question is whether the unlicensed ingestion of copyrighted material by generative-AI systems qualifies as a fair use.

In *New York Times* v. *OpenAI and Microsoft*[8], the New York Times alleged that OpenAI utilised millions of its articles to train its AI models. It argued that OpenAI's generative AI tools could potentially reuse its reporting and present that content on platforms like ChatGPT. Several prominent newspapers allied with the New York Times, alleging that OpenAI misappropriated their reporters' work to develop its generative AI systems. In defence, OpenAI maintains that its models are trained exclusively on publicly accessible datasets, which could include copyrighted content. OpenAI asserted that its approach includes creating copies of the data for analysis, which it claims is protected u/s 107 of the U.S. Copyright Act of 1976, under fair use provisions, since these copies are not publicly accessible and are utilized exclusively for training its models. Additionally, OpenAI claimed that its training methods qualify as fair use and do not infringe upon any copyrighted material. To support its position, OpenAI cited precedent set in *Authors Guild, Inc.* v. *Google, Inc.*[9], where the U.S. Court of Appeals determined that Google's reproduction of entire books to create a searchable database of excerpts was considered fair use. In a separate lawsuit, in the

---

5 U.S. Const. art. I, § 8, cl. 8.

6 The Copyright Act, 1957, Sec. 52 (India).

7 Civic Chandran v. Ammini Amma, 1996 P.T.C. 16 (Ker.) (India).

8 New York Times v. OpenAI and Microsoft 1:23-cv-11195 (U.S. District Court, New York) filed on 27th December 2023.

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

Silverman v. *OpenAI Inc.*[10] *case,* comedian Sarah Silverman, along with authors Christopher Golden and Richard Kadrey, accused OpenAI of copyright infringement for using their books to create and distribute derivative works without their consent. OpenAI asserted that utilizing data for AI training and operation qualifies as fair use.[11]

4. Analytical Framework

In the United States, the jurisprudence on doctrine of fair usage is determined by four factors.[12] These include:

i. intent and the nature of use, whether for commercial or educational (non-profit) purpose;

ii. quality of copyrighted content;

iii. extent and importance of portion used in comparison to entire copyrighted work;

iv. impact of use on market and economic worth of copyrighted work.

Factor One: The first factor evaluates whether use of copyrighted work serves commercial purposes or non-profit educational purposes, and whether new work 'transforms' the original and 'introduces something new' to it. Generally, non-commercial use supports fair use, whereas commercial use weighs against it.

While commercial use is only one component of the first factor, it's not definitive by itself. Courts typically do not classify new work to be fair use if it infringes upon copyrighted materials for financial gains. But AI developers now claim a new defence called 'defence of transformative purpose.' The transformative purpose defence has emerged as crucial factor in infringement and fair use. AI developers contend that training generative AI on copyrighted materials is inherently transformative. These systems analyse the works to recognise the 'patterns inherent in the human-generated media.' Yet, courts emphasise that even in such uses, transformation must extend beyond mere reproduction intended for consumption. For instance, it should enable sharing of information regarding the underlying work or include details of work in a database. Courts have ruled that, while AI training is considered transformative use, it automatically does not

---

[10] Silverman v. OpenAI Inc., 3:23-cv-03416, (N.D. Cal.).

[11] Impact of AI on IPR (public comment), available at: <https://www.uspto.gov/sites/default/files/documents/Electronic%20Frontier%20Foundation_RFC-84-FR-58141.PDF> (last visited on 10th October 2024).

[12] § 107 - Fair use of copyrighted work, available at: <https://www.govinfo.gov/ /content/pkg/USCODE-2010-title17/pdf/USCODE-2010-title17-chap1-sec107.pdf > (last visited on 12th October 2024).

guarantee fair use. The Court in *Campbell* v. *Acuff Rose Music*[13] stated that fair use protection is enhanced when a work is "transformative."[14]

Factor Two: The second factor considers nature of copyrighted work to assess whether it's factual or creative. This factor is seldom decisive when underlying work is creative. As fair use requires an in-depth analysis, AI platforms trained mostly on factual works, are likely to support a fair use exception.[15] U.S. Supreme Court in *Andy Warhol* v. *Goldsmith*[16] determined that when a secondary work serves similar purpose like original, and is used for commercial gain, it becomes difficult to defend fair use.

Factor Three: The third fair-use factor examines how much copyrighted work can be used and its importance in relation to the entire work. When a user copies the entire work, or its core creative elements, it negatively impacts the fair use argument, particularly if multiple complete works are involved. Nonetheless, Courts have determined that it may be acceptable to copy an entire work, if doing so is essential to achieve a transformative purpose. But even in these cases, the user may not take more than what is necessary to achieve the targeted transformative purpose. However, it is uncertain as to how much material can be taken from a copyrighted work and can still qualify as fair use. For example, in the case of writing a review of a book, the Fair use might allow that one can take out paragraphs for the review. However, it is unclear how many paragraphs can be removed, before it is classified as derivative or becomes a new composition.

In numerous instances of AI generation and machine training, there is a practice of extensively copying multiple copyrighted works. This suggests that AI generators extract as much material as they can from expressive works, including key creative components, to effectively train and generate high-quality results. For instance, *Open AI* referenced the ruling in *Authors Guild* v. *Google*[17], arguing that the focus of the third factor isn't on the quantity of copyrighted material copied, but on how much of work is publicly accessible. Open AI recognized that utilizing entire works was 'reasonably necessary' for developing an accurate AI, but significant copying is irrelevant if that copy is not made public. Consequently, *OpenAI* clarified that training data is not publicly available, instead, only the content produced from it is shared OpenAI subsequently maintained that its program represents a transformative use and therefore falls within the scope of fair use. Yet, this reasoning remains tenuous under the Copyright Act, since adopting it could

---

13 Campbell v. Acuff Rose Music, 510 U.S 569 (1994).

14 *Campbell,* 510 at 569.

15 Guild v. Google Inc, 804 F. 3d 202 (2d Circuit, 2015).

16 Andy Warhol v. Goldsmith, 598 U.S. 508.

17 *Authors Guild*, 804 F.3d at 202.

compromise the reproduction rights of creators. Additionally, the fair use exception clearly guides courts to evaluate the quantity and significance of copyrighted work being used, instead of relying upon a judicially established 'public access theory'.

The claim by AI developers that their training copies are never made public is not convincing, as it remains uncertain whether and how repositories of unauthorised works created for training are protected from further distribution and reproduction. Therefore, although the third factor is not conclusive, it largely depends on each situation, when complete works generally count against a fair-use exemption.

Fair use of data is primarily assessed when the data is factual. However, it may also qualify as fair use if the work has a creative aspect.[18] Nonetheless, creative works can be utilised to train AI with factual data, suggesting that the data components used for AI training are considered factual in nature. The previous proposition was to discourage the use of entire copyrighted work. However, with the growth of technological innovations, this position has changed. Now entire usage may be allowed for a transformative purpose.[19] However, it is crucial that the intent of AI developer aligns with content owners and both focus on distinguishing elements that users are looking for.

Factor Four**:** The fourth factor in fair use analysis examines the impact of infringing use on the value of copyrighted work. This factor argues against fair use, if the infringing work serves as a substitute for copyrighted work, particularly when it affects the markets where the copyright owner is active. *OpenA*I contended that since the dataset is processed by machines and not directly by humans, authors would not risk losing market or audience. However, the New York Times sued *OpenAI*, claiming that the AI tool 'substantially diminished the need for users to visit the publisher's website.'[20]

Overall, using copyrighted materials for AI training could negatively impact the market and value of original works. However, AI developers are reluctant to provide compensation to copyright owners for utilising their works in training generative AI. This happens even though numerous copyright owners are willing to provide licenses for AI training. So, the use of copyrighted work without a license also destroys the copyright owners' licensing market. In other words, offers of licenses indicate a fact that even copyright owners want to participate in the development of artificial intelligence.

---

18 Stewart v. Abend, 495 U.S 207(1990).

19 Kelly v. Ariba Soft, 280 F.3d 934.

20 B. Allyn, *'NYT' considers lawsuit against OpenAI as copyright tensions swirl*, NPR (16th August 2023), *available at* <https://www.npr.org/2023/08/16/1194202562/new-york-times-considers-legal-action-against-openai-as-copyright-tensions-swirl> (last visited on Oct. 15, 2024).

The value of copyrighted work should not deteriorate after its use in AI Training. This implies that AI should not serve as a substitute for original content by excessively relying upon it for training. Although, indeed, owning a copyright does not guarantee that the owner will receive all profits from his/her work, copyright does guarantee protection against substantial losses resulting from unauthorised use of their copyrighted material.

5. Burden of Proof and Substantially Similar Output - Two Additional Factors

The US Courts, in addition to the above four factors, rely upon two more tests to determine whether there is any infringement of copyright. The first criterion requires the plaintiff to show that the software had 'access to their works' and provide evidence of actual copying of original material. The second criterion is that the software must generate a 'Substantially Similar Output.' However, second criterion can be challenging to assess because it involves multiple factors, including the 'similar concept and feel', 'overall look and feel' and the 'inability of an average person to distinguish between both works.' Consequently, the determination is subjective.[21]

Thus, doctrine of fair use depends upon the facts of a particular case. Using copyrighted works for AI training does not fulfil requirements for fair use. The Fair use doctrine has usually given unexpected results on the application of four factors, when applied in the context of new AI technologies. For instance: In a case involving *Sega*,[22] U.S. Court concluded that Accolade's reverse engineering of Sega's 'Genesis video game software' qualified as fair use, despite involving copying copyrighted code. The court determined that first factor of fair use doctrine, concerning 'purpose and character of use,' supported Accolade. This was based on Accolade's objective to develop Genesis-compatible games for both 'legitimate and non-exploitative purpose.' Accolade's replicated protected code to determine the functional requirements for ensuring compatibility with Genesis console. Regarding the second factor i.e., the 'nature of copyrighted work,' the court stated that Sega's video game software received less protection than conventional literary works. This was attributed to inclusion of unprotected functional elements, including compatibility with Genesis console. Therefore, court decided in the *Sega* case that Accolade's reverse engineering qualified as fair use.[23]

---

21 *Generative AI & Copyright laws*, CRS (29th September 2023), *available at:* <https://crsreports.congress.gov/product/pdf/LSB/LSB10922> (last visited on 18th October 2024).

22 Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1527 (9th Cir. 1992).

23 *Id.*

6. The Doctrine of Transformative Use

In fair use doctrine, the first factor referring to 'the purpose and character of the use, including whether it's commercial or for non-profit educational purpose' has taken precedence over second and third factors due to emergence of 'doctrine of transformative use.'[24] The first factor in fair use analysis examines whether the use is transformative. This involves assessing whether the original work has been altered to create something new, such as fresh insights, aesthetics, or understandings.[25] The 'transformative use analysis' has been applied in cases where internet search engines use text and images for functioning. The concept of transformative use has become the fulcrum of modern fair-use analysis. The Supreme Court's decision in *Campbell v. Acuff-Rose Music, Inc.* marked a doctrinal shift from strict reproduction analysis to an inquiry into whether a secondary use adds new expression, meaning, or message.[26] The Court emphasised that transformation, rather than the commercial or non-profit character of a use, should carry the greatest weight under the first statutory factor. Search-engine cases extended this reasoning to functional copying.

The Court in *Perfect 10 Inc.* v. *Amazon.com Inc.*,[27] ruled that showing copyrighted images as thumbnails in search engine results is a fair use, highlighting its transformative character. While original images were intended for 'entertainment, aesthetic, or informative purpose,' the search engine repurposed them into a 'pointer' that guided users to source of content.[28] Similarly, the Court in *Authors Guild Inc.* v. *Google Inc.*,[29] determined that Google's digitisation of copyrighted books was fair use, emphasizing the transformative role of Google Books search platform. The court determined that, despite Google scanning entire copyrighted texts, it only showed snippets that served as pointer, guiding users to a wide range of books. This established a precedent in search engine cases for upholding a fair use defence.[30]

If internet search engines are deemed transformative, it paves way for other AI applications to be recognized as transformative in their use of expressive data. Similar rulings apply in cases where images, text, and videos are utilized as input to train models, enabling machines to produce

---

[24] J. Ginsburg, *Fair Use in U.S.: Deformed, Transformed, and Reformed?* SJLS 265-94(2020).

[25] P. Leval, *Towards the Fair-Use Standards*, 103 **HAR. LAW REV. 1105 (1990).**

[26] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994).

[27] Perfect 10 Inc. v. Amazon.com Inc., 508 F.3d 1146, 1167 (9th Cir. 2007).

[28] *Id.*

[29] Authors Guild Inc. v. Google Inc., 804 F.3d 202, 216 (2d Cir. 2015).

[30] *Authors Guild*, 804 at 216.

creations comparable to humans' creativity. Machines have the ability to understand and generate new images and insights. Furthermore, large-scale reproduction of images or text processed by computers are not restricted, as affirmed in *Author's Guild Inc.* v. *Google Inc.* In this case, Google digitised millions of books for its search engine. The Court's decision highlighted that transformative nature of a product can take precedence over its commercial purpose.

However, Courts have many times fluctuated in holding fair use doctrine in favour of AI technology.[31] In the *Zillow Grp. Inc.*[32] case, the court noted that Zillow's use of copyrighted photographs in its apartment listings did not qualify as fair use. The platform featured images of stylish rooms, which users could filter based on criteria such as colour, price, and room type. Although the search engine displayed only those photographs that were searchable by function, the court ruled that this did not alter their original intent to artistically showcase properties and rooms, thus maintaining the character of original photos. In addition, the plaintiff was seeking opportunities to license its photographs, which made the fourth factor i.e., 'the impact of use on potential market or value of copyrighted work' favourable to the plaintiff and unfavourable to the defendant, Zillow.[33] In a similar case, the court in *Fox News Network LLC v. TVEyes Inc.*[34] determined that despite the transformative character of the TV clip search engine, the use was not fair, as it encroached upon Fox's potential licensing market.[35]

As a result, rulings in *VHT, Inc.*[36] and *TVEyes*[37] demonstrated that fair use defence is complex and not easily applicable in the context of AI technologies. While AI developers may invoke the first factor by highlighting the transformative nature of their use, the cases of *VHT Inc.* and *TVEyes* demonstrate that courts often give considerable weight to the potential impact of the end use on licensing markets. Consequently, an AI developer could face penalties under fair use doctrine for preventing a copyright owner from licensing their work for inclusion in training datasets.

---

31 Michael W. Carroll, *Copyright & Progress of Science-Why TDM is Lawful*, 53 U.C. Davis Law Review, 893-963 (2019), *available at* <https://lawreview.law.ucdavis.edu/sites/g/files/dgvnsk15026/files/media/documents/53-2_Carroll.pdf> (last visited on Oct. 20, 2024).

32 VHT Inc. v. Zillow Grp Inc., 918 F.3d 723 (9th Cir. 2019).

33 *Id.*

34 Fox News Network LLC v. TVEyes Inc., 883 F.3d 169, 178–80 (2d Cir. 2018).

35 *VHT*, 918 F.3d at 723.

36 *Id.*

37 Fox News Network LLC v. TVEyes Inc., 883 F.3d 169, 178–80 (2d Cir. 2018).

It means that copyrighted photographs do not get 'weaker protection' even if they are used for transformative or informational purposes.[38] In *Andy Warhol Foundation for the Visual Arts, Inc.* v *Goldsmith*,[39] the Court held that Andy Warhol's artworks, created from the photograph of the renowned American singer, songwriter, and producer did not meet the criteria for being transformative. The Court explained that introducing a new aesthetic or expression to the original work alone is insufficient to qualify as transformative. The Court determined that for a work to be deemed transformative, it must serve a clear artistic intent and significantly modify the original work.

AI copyright ownership in a non-human entity-generated work is also complex. In *Naruto* v. *Slater*,[40] also known as '*monkey selfie*' case, the United States Court explained the meaning of personhood in copyright law. In this case, wildlife photographer David Slater accidently left his camera unattended, allowing a macaque to pick it up and take selfies of itself, which later gained fame as 'monkey selfies.' Despite macaque taking photographs, Slater credited himself for them. An animal rights group called 'People for the Ethical Treatment of Animals' opposed Slater's claim and made claim for the monkey as the original copyright holder. However, the U.S. court determined that the monkey lacked legal standing for making such a claim, as U.S. Copyright Act does not explicitly allow animals to initiate copyright infringement lawsuits. Thus, by making a parallel between two non-human entities whether it is monkey or the AI, any legal framework may or may not explicitly recognise the right of a non-human entity, including machines, to own a copyright or sue for copyright infringement.

The '*monkey selfie*' judgment initiated on the need to develop copyright law suitably applied to new technologies and AI-generated content, because as per the current copyright law, only works of human creation is eligible for copyright protection. This indicates that creations produced by artificial intelligence do not qualify as works of authorship under copyright law. Even when an AI developer justifies their use of copyrighted photographs by invoking an idea-expression dichotomy, claiming that they used the informational content for machine training - the Courts may conclude that the developer did not change the original purpose of the work in assembling the training dataset.

These inconsistent outcomes reveal that transformation remains a fact-sensitive and unpredictable inquiry. In the context of AI, developers contend that machine-learning systems analyse works to extract patterns rather than

---

38 Brammer v. Violent Hues Prods. LLC, 922 F.3d 255.

39 Andy Warhol Foundation for the Visual Arts, Inc. v Goldsmith, 598 U.S. 508 (18th May 2023).

40 *Naruto v. Slater*, No. 16-15469 (9th Cir. 2018).

to reproduce them, thereby creating a new, non-expressive purpose. Yet the lack of a consistent judicial standard triggers the question whether large-scale copying for training generative models constitutes transformation. Scholars such as Nimmer argued that 'transformative use' has drifted from its parodic roots into an overbroad justification for commercial reproduction, risking doctrinal incoherence.[41] As AI systems expand, courts will need to articulate whether data ingestion for pattern recognition truly adds something new or merely repurposes expressive content in bulk.

Copyright law must evolve to strike a balance between a country's economic interests and the moral rights of creators. Under current law, only human authors are granted specific rights, including economic rights to monetise their work, as well as moral rights like attribution and integrity. But there is a need to look at the rights for AI-generated works. It is imperative that copyright law adapt to grant AI developers a clearer statutory defence for incorporating expressive copyrighted materials into training datasets, thereby reducing dependence on the uncertain contours of fair use.

7. India's Legal Framework for Using Copyrighted Material in AI Training

India's copyright law follows the doctrine of fair dealing, not fair use. In India, Section 52 of the Copyright Act, 1957 outlines specific exceptions to copyright infringement, including use for private study, criticism, review, and reporting of current events.[42] Determination of fair use is also a complex subject in countries including India. Unlike the open-ended U.S. framework, fair dealing is a closed list of permissible acts, leaving limited space for judicial creativity. Copyright infringement is based on legal principles, facts and circumstances of each case. Section 14 of the Copyright Act, 1957, r/w Section 51 of the Copyright Act, 1957, addresses various types of infringement. Section 14 deals with the meaning of copyright for literary, dramatic and musical works, including those which are created through the use of computer programs; whereas section 51 mentions situations where copyright could be infringed. The doctrine of fair use, which is codified under section 52, elaborates instances when certain acts would not constitute infringement, such as when it is used for personal use, research, education or for the purpose of critique.[43] The Kerala High Court, despite a lack of legal precedent, articulated a four-factor test in *Civic Chandran* v. *C. Ammini*

---

41 Melville B. Nimmer & David Nimmer, *Nimmer on Copyright* § 13.05[A] [1][b] (2024).

42 The Copyright Act, 1957, § 52 (India).

43 Shlok Sharma, *Generative AI & Copyright* Conundrum, LEAFLET (May 16, 2023), *available at:* <https://theleaflet.in/generative-ai-and-the-copyright-conundrum/ > (last visited on Oct. 21, 2024).

Amma,[44] resembling the U.S. model, focusing on the purpose of the use, the nature of the work, the amount used, and the effect on the market.[45] However, Indian courts rarely invoke 'transformative purpose' as an independent test.

In *India TV Independent News Service (Pvt.) Ltd.* v. *Yashraj Films (Pvt.) Ltd.*,[46] the Delhi High Court held that unauthorised use of film clips for television broadcasting did not constitute fair dealing as it was not sufficiently connected to purposes of reporting or criticism.

This indicates that in India, the use of copyrighted data for AI training for non-commercial purposes is permitted. However, AI companies argue that all creations using generative AI should be classified under fair-use, as they modify original works by introducing new expression, meanings or messages. In response, original content owners contend that AI may inadvertently generate art or code that closely mirrors the original work, as such, it may fail to meet the 'transformative work' criteria, thereby disqualifying it from fair use. The government maintains that the present legal framework doctrine for fair-use, along with patent and copyright law, is adequately designed to safeguard AI and its related innovations.[47] This absence of transformative analysis leaves India ill-equipped to evaluate AI-training uses. The reproduction of vast datasets by AI systems could fall outside any enumerated fair-dealing exception, exposing developers to liability even for non-commercial research. Given the increasing importance of machine learning to innovation, Indian policymakers must consider a limited statutory exception permitting text and data mining (TDM) for lawful purposes, akin to reforms in other jurisdictions like the USA.

Furthermore, Indian law should clarify authorship for AI-assisted works. Currently, Section 2(d)(vi) of the Indian Copyright Act, 1957 deems the 'person who causes the work to be created' as the author for computer-generated works.[48] Courts have yet to interpret this provision in the AI context. Without guidance, uncertainty persists over whether developers, users, or the AI itself owns the resulting work. To foster innovation, India must modernise its fair-dealing doctrine by integrating transformative reasoning, balancing user rights with creators' interests and introduce clear statutory protection for bona fide data-mining activities.

---

44 Civic Chandran v. C. Ammini Amma, 1996 16 PTC 329 (Kerala) (India).

45 Id.

46 (India TV Independent News Service (Pvt.) Ltd. v. Yashraj Films (Pvt.) Ltd.), FAO (OS) 583/2011 Dt. 21-08-2012, 2012 SCC OnLine Del 5581 (India).

47 *Id.*

48 The Copyright Act, 1957, Sec. 2(d)(vi) (India).

8. Status of Data Mining In Copyright Law

Text and data mining (TDM) refers to an automated computational analysis of digital content to identify patterns, trends, and correlations, such as data, audio, images or other media, aimed at uncovering new insights.[49] AI systems rely on TDM to extract insights from vast textual or visual datasets. However, as TDM typically involves copying source materials into training corpora, it raises concerns under copyright law. Legal commentators contended that data mining cannot be treated as copyright infringement, rather it is lawful copying of unprotected material which should be extended to copyright protected material in the case of fair use. Fair use defence doctrine must also provide legal certainty to AI innovators because without it they may deter because of concerns related to infringement, prolonged litigation and steep fines. In *Feist Publications, Inc*. v. *Rural Tel. Serv. Co.*,[50] it was held that all copying is not copyright infringement. In *Baker* v. *Selden,* the court considered whether every use of a work's physical form qualifies as a copyright violation. Selden held the copyright to his book that outlined a new accounting system. Baker adopted a similar accounting system but used Selden's forms to explain the process.[51] These forms not only explained the accounting system but could also be used to perform accounting. Copyright law protects use of these forms for explanatory purposes, but it does not cover the forms as inventions, since inventions are governed by patent law. The U.S. Supreme Court decided that Baker was not liable for copyright infringement because he used Selden's forms as part of a new accounting system, rather than as a means to explain the system. The Court ruled that copyright infringement involves not only copying a work's material form, but also using it for its 'expressive purpose'. As a result, purely technical or non-expressive uses of a work do not constitute copyright infringement. It means that technical and non-communicative uses should not even be a subject of fair use analysis. Therefore, activities like data mining or downloading images from the internet for training AI models are not copyright infringement. Instead, AI developers don't publicly communicate the copyrighted images, rather use them for training a machine learning models. As copyright protection does not cover material forms of works, downloading images does not violate protected use of copyrighted material.[52]

---

[49] Jean-Paul Triaille, Jerome De Meeus, *Study of Legal Framework of TDM,* EC (March, 2014), *available at*: <https://op.europa.eu/nl/publication-detail/-/publication/074ddf78-01e9-4a1d-9895-65290705e2a5/language-en> (last visited on Oct 28, 2024).

[50] Feist Publications, Inc. v. Rural Tel. Serv. Co.499 U.S. 340, 361 (1991).

[51] Baker v. Selden, 101 U.S. 99 (1879).

[52] *Id*.

Thus, copyright law was designed to safeguard creative and expressive works. The idea-expression dichotomy reinforces this by limiting protection to only expressive elements of a work, excluding the functional ideas it may include.[53] An AI developer, while developing a machine learning training set, does not want to reproduce the expressive work. His main interest lies in the functional aspects, embedded within the material forms. For instance, developers of *Natural Language Processing* use literary works purely as training data to identify fundamental patterns in human speech. Therefore, holding AI developers liable for copyright infringement for using expressive works to train functional models would be an unjustified expansion of creators' rights.[54]

AI developers use images and videos of streets to train machines to detect pedestrians in computer vision systems.[55] In these instances, the machine is not duplicating the expressive elements of work, since the expression is inherent in writing or in artistic portrayal of streets in a photograph. In the *Authors Guild* case, Judge Level emphasised that the object of copyright is to promote public knowledge, with the public being the beneficiary.[56] Allowing data mining in copyright law would provide benefits to the public by stimulating innovation. Hindering the growth of beneficial AI technology would ultimately impede the goals of intellectual property law.[57]

9. Comparative Developments

Establishing a well-defined legal framework for text and data mining within copyright law would offer clarity alongside the fair use doctrine. The growing advantages of AI technologies have driven countries to revise their copyright laws, encouraging innovation and ensuring global competitiveness in the fields of AI and Machine Learning. Many jurisdictions have enacted text and data mining (TDM) exceptions in their copyright laws. For example:

Japan revised its copyright laws to introduce an exception for TDM.[58] Under Article 47(7) of The Copyright Act, 1970 of Japan, law permits TDM

---

53 Mazer v. Stein, 347 U.S. 201 (1954).

54 A.C. Mendes & C. Antunes, *Pattern Mining with Natural Language Processing*, IN **MACHINE LEARNING & DATA MINING IN PATTERN RECOGNITION,** Petra Perner (ed.) (2009).

55 A. Brunetti, et.al., *Computer Vision & Deep-Learning Techniques in pedestrian detection*, 300 Neurocomputing 17 (2018).

56 *supra* n. 58, 804 F.3d 202.

57 John Cormick, *AI Will Lower Risk of Future Wildfires*, WSJ (Sept. 11, 2020), *available at:* <https://www.wsj.com/articles/california-utilities-hope-drones-ai-will-lower-risk-of-future-wildfires-11599816601> (last visited on Oct. 24, 2024).

58 *Japan to Meet Future Demand in AI & Big Data*, EARE (3rd September, 2018), *available at:* <https://eare.eu/japan-amends-its-copyright-legislation-to-meet-future-dem ands-in-ai-and-big-data/> (last visited on Oct 29, 2024).

for all users, regardless of whether commercial or non-commercial. The Act included three provisions aimed at clarifying law and addressing prior copyright restrictions on AI. Article 30-4 of the Act permits users to interpret copyrighted works for machine learning purposes. Article 47-4 authorises reproduction of works in digital form, while Article 47-5 enables use of copyrighted content for verifying data.[59]

The European Union's Directive 2019/790/EU introduced 2 exceptions for TDM. Specifically, Article 3 of the EU directive 2019 allows research organisations to reproduce content for scientific research, provided they have lawful access to the material. Article 4 of the Act creates an exception for reproducing lawfully accessible works for commercial TDM. It gives rights holders a choice to opt out of this exemption. While the TDM exceptions in the Directive recognise the value of data mining for research and technological advancement, restrictions outlined in Article 4 put commercial AI developers at a disadvantage.[60] The opt-out mechanism under EU law demonstrates ongoing tension between innovation and control. While it preserves authors' autonomy, it imposes transaction costs that disadvantage small developers. As the World Intellectual Property Organisation (WIPO) has noted, jurisdictions lacking clear TDM exceptions risk deterring AI research and global competitiveness.[61]

The United Kingdom similarly codified an exception for TDM for research under its copyright framework. Under § 29A of its Copyright, Designs and Patent Act, 1988, copies created for TDM are not a copyright violation, as long as the activity is conducted exclusively for research purposes. In the U.K., the exception is limited to those with legal access to work and doesn't apply to start-ups or entrepreneurs seeking to develop innovative machine learning technologies for commercial use.[62]

Australia, Canada and Singapore also recognise TDM exceptions that balance innovation with authorial rights. While in the United States, data mining permits reproduction, creation of derivative works, and sharing of datasets for TDM. It permits both commercial and non-commercial uses, though restricted to functional, non-expressive purposes of TDM.[63]

India's current framework lacks an explicit TDM exception. Consequently, every reproduction for AI training must rely on case-specific fair-dealing defences, creating legal uncertainty. To align with global best practices, India could adopt a statutory TDM exception modelled on the

---

59 Copyright Act of Japan, art. 47-7 (as amended 2018).

60 Directive (EU) 2019/790, arts. 3–4, 2019 O.J. (L 130) 92.

61 WIPO, Study on Exceptions and Limitations for Text and Data Mining (2020).

62 Copyright, Designs and Patents Act 1988, § 29A (UK).

63 supra n. 59

Japanese approach, permitting both commercial and research uses when the copying is for non-expressive analytical purposes.

10. Fixation of Legal Uncertainty and Copyright Dilemma

10.1. Providing Legal Certainty to AI innovators

Data mining falls outside the scope of copyright law. Extending the fair use doctrine over it will eliminate any bias against AI within copyright law. This approach is crucial because fair use has historically been referred to as 'the most complex area of copyright law.'[64] These changes are feasible because fair use is a dynamic doctrine, allowing adaptability to new contexts, including the use of copyrighted works in AI applications. Moreover, at a time when data mining is not considered as a copyright infringement, it will definitely not fall under judicial scrutiny under the fair use analysis. Such certainty in copyright law is essential because the possibility of infringement litigation has already deterred many Start-up creators.

Downloading data from the internet for training purposes does not qualify as using protected works. Fair use is designed to protect specific communicative uses of copyrighted material, including parody, art, and criticism. An AI developer is not accountable for using the work's material form, as opposed to the expressive or communicative elements. Examples of non-infringing uses include search engine thumbnails and temporary copies created during web-browsing.

Those supporting the fair use doctrine in data mining should recognise that, although this doctrine is traditionally applied to music, art, and literature, extending it to other creative fields could lead to negative consequences. The U.S. Supreme Court in *Campbell* v. *Acuff-Rose Music Inc*. stated that a musical parody was not likely to replace an original song, since both works serve different purposes in the marketplace.[65] In *Cariou* v. *Prince*,[66] two artists targeted different segments of art market. The plaintiff earned a modest income from royalties by selling his artwork mostly to personal contacts, whereas the defendant made millions by selling his art to high-profile celebrities. The U.S. Supreme Court, by expanding the logic of works serving different market functions, held that differences in wealth justify a finding of fair use. This argument implies that extending the fair use doctrine to digital technologies could restrict authors' rights in various other domains.

---

64 VHT Inc. v. Zillow Grp. Inc., 918 F.3d 723.

65 *supra* n. 25, 510 U.S. 569 (1994).

66 Cariou v. Prince, 714 F.3d 694.

The 'transformative use analysis' under fair use doctrine offers a new framework for courts, guided by the 'principle of stare decisis.' The broadening of transformative use is contributing to increased uncertainty in copyright law. For instance, the Court in Tiffany Design Inc. v. Reno-Tahoe Speciality Inc., determined that intermediate copying constitutes copyright infringement.[67] Thus, there have been conflicting judicial decisions on fair use. AI developers are confused as to how their investment in AI technology will be viewed by the courts. So, AI innovators must be provided a statutory defence while developing new technologies?[68]

10.2. Handicaps new AI Developers Disproportionately

Uncertainty in copyright law disproportionately handicaps small AI developers, as large AI tech companies have resources to access costly data and the best legal team to litigate, but small developers lack resources in both areas. These deterrents affect small developers in using large datasets. For instance, major technological platforms such as YouTube and Meta include service conditions that allow them to access copyrighted content uploaded to their servers.[69] When users upload content onto YouTube, the platform receives a globally royalty-free and non-exclusive license to use that content. This data serves as a valuable asset for training YouTube's machine learning models. Meta utilises data from its more than 2 billion users to advance its facial recognition systems and create text for the visually impaired.[70] Even when large companies lack internal systems for collecting data, they can still purchase vast datasets.

In contrast, small innovators do not have equivalent access to datasets.[71] Though open-source datasets exist, they may be prone to bias. However, if a statutory safe way is provided, that data mining of copyrighted content for machine learning training is legalised, which would enable smaller developers to enjoy free access to valuable data and develop new innovative

---

67 Tiffany Design Inc. v. Reno-Tahoe Speciality Inc., 55 F. Supp. 2d 1113 (1999).

68 Peter Ned, 'Comment to: *In the US, is it Illegal to Train Neural Networks Using Copyrighted Images?* QUORA (July 1, 2017), *available at*: <https://www.quora.com/In-the-US-is-it-illegal-to-train-neural-networks-using-copyrighted-images > (last visited on Oct 23, 2024).

69 See *Terms of Service*, YOUTUBE *available at:* <https://www.youtube.com/static?template=terms > (last visited on Oct 24, 2024).

70 Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 Washington. L. Rev. 579, 606–07 (2018), *available at*: <https://digitalcommons.law.uw.edu/cgi/viewcontent.cgi?article=5042&context=wlr > (last visited on Oct 25, 2024).

71 S. Levy, *Inside Facebook's AI Machine*, THE WIRED (23rd February 2017), *available at:* <https://www.wired.com/2017/02/inside-facebooks-ai-machine/> (last visited on Oct. 26, 2024).

products to compete against technology giants. Increased competition would ultimately benefit the public by providing access to a greater number of high-quality AI products.

Large tech companies and AI players are able to create unauthorised copies of protectable works to use as training data for AI systems. Even if such companies run contrary to copyright law in training their machine learning algorithms, they are better equipped with expensive lawyers to defend themselves from liability. As the doctrine of fair use is highly fact-specific and unpredictable, it creates space for smart, creative, and expensive lawyers. For example, Google secured a victory in fair use case of *Authors Guild* v. *Google, Inc.*,[72] even after 10 years long litigation.[73] But in the case of small developers, litigation can drain the bank accounts and cause the closure of their operations. Support of powerful institutions as in *Cariou* v. *Prince*[74] proves influence of money in securing favourable verdicts.[75] Thus, present regime of copyright law is causing harmful deterrent effects for new AI developers. Legal uncertainty and threat of exorbitant fines also deter smaller actors from creating and using valuable datasets.

10.3. Licensing of Author's Rights Can Propagate Bias

Experts suggest the use of licenses if AI training data is used in commercial applications. Requiring licenses for training data could impose substantial challenges and overstep authors' rights. Typically, data for machine learning models is collected through automated web-scraping to meet the enormous demand for data.[76] Additionally, reproductions involved during data mining do not constitute infringement. In *Baker* v. *Selden*, the Court determined that an author cannot assert copyright over every reproduction of their work's physical form. Copyright is restricted to the expression of ideas in a communicative sense, not the material representation of those ideas. Experts note that AI developers have adapted to uncertainty surrounding copyright by using datasets released under Creative Commons (CC) licenses,[77] or

---

72 *supra* n. 28, 804 F.3d 202 (2d Cir. 2015).

73 A. Liptak & A. Alter, *Challenges to Google Books Denied by Supreme Court*, NYT (18th April, 2016), *available at:* <https://www.nytimes.com/2016/04/19/technology/google-books-case.html> (last visited on Oct. 26, 2024).

74 *Cariou v. Prince*, 714 F.3d 694 (2d Cir. 2013).

75 D.W. Kenyon & S.P. Demm, *SC Denies Certiorari in Cariou Fair-Use Case: What's Next?* IP MAGAZINE (February 2014), 71- 72, *available at:* <https://www.huntonak.com/media/publication/3211_IPMagazine_CariouArticle_2014.pdf> (last visited on Oct. 27, 2024).

76 C. Hansen, *Web Scraping for the Machine Learning along with SQL Database,* ML (Dec. 4, 2019).

77 *Updated-Dataset*, YouTube8M, *available at*: https://research.google.com/youtube8m/download.html> (last visited on Oct 27, 2024).

by publicly accessing text corpora from platforms like Wikipedia, which permits free access, copying, and redistribution.[78]

In *Andersen* v. *Stability AI Ltd.*,[79] 3 artists sued several generative AI platforms, accusing them of using their copyrighted images without authorisation to train AI models for generating images. The Court ruled that works created by AI are derivative, as they are essentially reproductions of original copyrighted images. In *Author's Guild* v. *Google*, Google digitised printed books to develop Google Books, prompting lawsuits from original authors. The court ruled that Google's action of digitising and storing books was a fair use. It evaluated the 'purpose factor' of fair use and found that Google's intent differed from that of authors, as books were not being used for their content. Instead, they were essential for creating a database, serving different purposes from their original. Therefore, the Court ruled it as fair use.

11. Conclusion and Suggestions

Machine learning should not be used as an excuse to break the law. That is why AI developers need to ensure that the data they use to train and improve algorithms is collected and stored responsibly.[80] Artificial Intelligence, in its initial stage, faced the limitation of adequate data and information, which resulted in producing factually incorrect information. Following extensive training on vast datasets, AI has become accurate, dependable, and user-friendly. Nevertheless, it continues to evolve, striving for improved accuracy, especially in social and technological areas. But in the legal sphere, several issues need to be resolved, and one such is copyright infringement.

AI content creation faces a threat from the fair use doctrine. The issue is whether the AI-generated work is a derivative work for commercial exploitation, or whether it is limited only to education, research, teaching, or other purposes. In practice, generating AI-based works require intricate steps before producing any output. Thus, AI-generated content may possess an inherently transformative quality, thereby falling within the scope of fair use, as it modifies the character of the source material.

---

78 *Wikipedia:Copyrights*, WIK1PEDIA, *available at*: <https://en.wikipedia.org/wiki/Wikipedia:Copyrights> (last visited on Oct 27, 2024).

79 *Getty Images (US) Inc.* v *Stability AI Inc.*, WHO, *available at:* <https://docs.justia.com/cases/federal/district-courts/california/candce/3:2023cv00201/407208/67> (last visited on Oct 29, 2024).

80 D. Coldewey, *Amazon Inc. Settles with FTC for $25M After 'Flouting' Kids'* Privacy, TECH CRUNCH (31st May, 2023), available at: <https://techcrunch.com https://techcrunch.com/2023/05/31/amazon-settles-with-ftc-for-25m-after-flouting-kids-privacy-and-deletion-requests/> (last visited on Oct 29, 2024).

The legal framework should address data mining within the context of machine learning and AI, emphasising the automated process of extracting functional insights from expressive data. A broad definition of data employs other techniques for pattern extraction. The legal framework should also be specific on the rights conferred to data miners. To provide legal clarity, it should also be beneficial to smaller innovators. The legal framework should clearly specify that reproductions for labelling and annotating works are permissible within data mining. Furthermore, the system should establish regulators to monitor data usage, ensuring fair, safe, and high-quality AI products. Commercial uses should not be subject to restrictions like opt-out mechanisms, as they impede innovation and contradict copyright law. They should be permitted to the same degree as research and non-commercial uses.[81]

Copyright holders should not have the option to opt out of allowing their works to be used for commercial data mining. This could create a licensing market for machine learning data, creating legal ambiguity. Such move would disadvantage smaller innovators, potentially leading to low-quality and biased AI systems. The legal framework emphasises the functional, non-expressive uses of data mining, as expressive AI-generated works could compete with and even replace original creations, placing a strain on authors and creators.

Artificial intelligence challenges the human-centred premises of copyright law. The doctrines of fair use and fair dealing, though conceptually similar, diverge in flexibility and adaptability. The U.S. model's emphasis on transformation enables functional uses like search indexing and TDM, but its unpredictability creates litigation risks. India's closed-list approach ensures certainty, yet constrains innovation. To reconcile these tensions, the following reforms are proposed:

1. Statutory Clarification of Authorship: Legislatures should define authorship for AI-assisted works, recognising human oversight while preventing the exclusion of machine-generated creativity.[82]
2. Explicit TDM Exception: Adopt a statutory exception for TDM, covering both commercial and non-commercial uses, modelled on Japan's Article 47-7 and the EU Directive's dual-tier approach.[83]
3. Safe Harbour for Developers: Provide legal protection for bona fide AI training that uses lawfully accessed materials for non-expressive analysis.[84]

---

[81] C. Geiger, et.al. *TDM: Articles 3, 4 of Directive - 2019/790/EU* 36 (Center for IP Studies) 2019.

[82] The Copyright Act, 1957, § 2(d)(vi) (India).

[83] Copyright Act of Japan, art. 47-7; Directive (EU) 2019/790, arts. 3–4.

[84] 17 U.S.C. § 107 (2018).

4. Balanced Licensing Mechanisms: Encourage voluntary collective licensing to compensate creators without imposing prohibitive transaction costs.[85]
5. Integration of Transformative Analysis into Fair Dealing: Indian courts should interpret Section 52 dynamically to incorporate transformative reasoning, ensuring flexibility without legislative overhaul.[86]

These measures will help in aligning copyright law with the realities of machine creativity, ensuring that legal doctrine promotes, rather than restrains technological progress.

[85] OECD, *AI, Copyright and Innovation: Policy Perspectives* (2023).

[86] Civic Chandran v. Ammini Amma, 1996 P.T.C. 16 (Ker.) (India).