# THE (NEW) IMITATION GAME: EXAMINING COPYRIGHT CLAIMS IN TRAINING DATA USED IN VISUAL GENERATIVE MODELS

***- Vineet Jadhav****

## ABSTRACT

*Large language models have found their way into the mainstream of daily life. These models require rigorous training and development using literary and other artistic works created by humans as a reference point. This creates a tension between scientific and technological progress, and the rights of exclusive exploitation vested in the authors of these works by statute. With the rise of large language models, both in their use and in the development of novel ones, developers have been found to use copyrighted material to train and develop these AI models, leading to a rise in copyright-focused lawsuits in courts. This paper aims to examine copyright infringement claims that can arise as regards the use of copyrighted works in training visual generative models, which include chatbots and programs which generate images based on reference text or images. The paper examines the manner in which such models are trained, which serves as a primer for the technology involved. Different theoretical perspectives that arise as a result of the technical understanding of generative models are then discussed insofar as they contribute to answering questions of infringement and regulation. The paper also discusses several contemporary lawsuits filed and pending before courts which involve infringement claims against developers of generative models arising out of the use of copyrighted works to train their generative models. The paper concludes that claims under Indian law, maybe maintained on grounds of the rights to reproduction, communication to public, and an interpretation of fair dealing. Finally, suggestions are made to balance the rights of authors with the contemporary developments in Generative Artificial Intelligence.*

**Keywords:** Copyright infringement, Large Language Models, Generative Artificial Intelligence, Fair Use, Right to Reproduction.

* Student, Hidayatullah National Law University.

## INTRODUCTION

The last demi-decade has seen an exponential rise in the capabilities of generative artificial intelligence. Fuelled by the rave success of models such as Google's Gemini and OpenAI's ChatGPT, various companies have made developing proprietary generative models a priority. These models require extensive training and development, the bulk of which is based on extant copyrighted work, such as articles, art, photographs, web-pages, etc. This data forms the basic building blocks to "teach" the generative model so it can generate novel material on its own. The simple question that this paper will try to answer is — Does the use of copyrighted material in the training of an image-generating artificial intelligence model (hereinafter 'generative model') give rise to any claims under copyright law? Copyright legislations across the world do not comprehensively or expressly deal with the unique issues posed by artificial intelligence, and the impact of using works to train these models. As such, various claims can be made based on existing provisions of the law. This paper aims to examine these claims as presented in select lawsuits, and under Indian law. This paper will be divided into several sections — starting with a primer on the manner in which visual generative models create images, it moves to Part III, which examines the claims in recent lawsuits made against developers of various generative models. Part IV examines theoretical perspectives on training data and infringement, and Part V aims to apply Indian law to the issue and examine the current status of the law; Part VI suggests recommendations to remedy the various issues that are identified in the course of the paper, and Part VII concludes the paper. The paper is limited to the extent that empirical data on the exact mechanism of training various models is not made publicly available, or is undergoing discovery in various lawsuits.

## VISUAL GENERATIVE MODELS: A PRIMER

Artificial Intelligence broadly means the ability of computers to process information the way human beings do.[1] It entails a machine's capability to perform intelligent tasks such as adapting to situations, responding to conversations, being capable of rationality, having goal-driven behaviour, being aware of oneself and generating novel content.[2] Developers deploy various methods to develop artificial intelligence, but the one process that is common to all is machine

[1] B J Copeland, 'Artificial Intelligence' (*Britannica*) <www.britannica.com/technology/artificial-intelligence> accessed 15 February 2024.
[2] Roger C Schank, 'What Is AI, Anyway?' (1987) 8(4) AI Magazine 59 <www.doi.org/10.1609/aimag.v8i4.623> accessed 8 July 2025.

learning.[3] Machine learning is the process of training a machine to understand a certain task based on input and output data, and to create the output from a given input. This process is undertaken to gradually increase the accuracy of the model's predictive capability, so that it can generate accurate results on being asked to perform a function.[4] A distinctive trait of machine learning lies in the fact that the model has an inherent capacity to gain "experience" to build knowledge about a task on its own.[5]

There are various methods by which machines learn. Machine learning entails observing the input and output passing a neural network, and then ascertaining the process by which the output was generated by the network.[6] A neural network can be understood as a biomimetic concept, consisting of "neurons" which, like neurons in the human brain, are processing units connected with each other, and through which data transfers take place and successive iterations of the intended output are created.[7] There are multiple such layers of neurons, which, in tandem with the weights and bias, create successive iterations of an output to find an optimal balance of the weights to create the intended output.[8] Weights can be understood as numerical values which determine the relative importance and strength of various learned characteristics of the intended output. The weights either promote or discourage the prominence and frequency of any given characteristic.[9] The optimal balance of these weights is regulated and measured in the network by another mathematical value called "bias" which measures how far-off the current iteration of the output is from the intended result.[10] This process of creating outputs is repeated for a copious number of times, and the model is incentivised to find the correct balance between the weights by promoting reduction in bias, and by penalising it, incorporating a mathematical value called "reconstruction loss", which tells the model to recalibrate its

---

[3] Niklas Kühl and others, 'Machine Learning in Artificial Intelligence: Towards a Common Understanding' (*Hawaii International Conference on System Sciences*, 2019) <www.arxiv.org/abs/2004.04686> accessed 15 February 2024.

[4] 'What is Machine Learning?' (*IBM*, 22 September 2021) <www.ibm.com/topics/machine-learning> accessed 15 August 2023.

[5] Sara Brown, 'Machine Learning, Explained' (*Ideas Made to Matter*, 21 April 2021) <www.mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained> accessed 15 February 2024.

[6] Anders Krogh, 'What Are Artificial Neural Networks?' (2008) 26 Nat Biotechnol 195 <www.nature.com/articles/nbt1386> accessed 8 July 2025.

[7] Laurent Pujo-Menjouet and Clement Viricel, 'Algorithms: A Biomimetic Approach to Performance and Nuance' (*Polytechnique Insights*, 25 October 2023) <www.polytechnique-insights.com/en/columns/science/algorithms-a-biomimetic-approach-to-performance-and-nuance/> accessed 4 February 2024.

[8] Anders Krogh, 'What Are Artificial Neural Networks?' (2008) 26 Nat Biotechnol 195 <www.nature.com/articles/nbt1386> accessed 8 July 2025.

[9] Osval Antonio Montesinos López, Abelardo Montesinos López and José Crossa, *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (Springer 2022) 381.

[10] Lukasz Gebel, 'Why we need Bias in Neural Networks' (*Towards Data Science*, 21 August 2020) <www.towardsdatascience.com/why-we-need-bias-in-neural-networks-db8f7e07cb98/> accessed 8 July 2025.

approach. This way, the model gains "experience" and understands the precise import of the intended output.[11] The specific way this optimal balance is found differs from the kind of neural network being used.[12]

Visual generative models are usually trained using Variational Autoencoding (hereinafter 'VAE') or Generative Adversarial Networks (hereinafter 'GAN').[13] A VAE is a model which consists of an encoder, which can be conceptualised as a learning-half, and a decoder, which is the generative-half.[14] The encoder receives an input, such as an image of a face, and it adds noise and other forms of abstraction to create a "bare minimum" of the image. The decoder then picks up this "bare minimum" image and identifies attributes such as face shape, gender, colour, features like eyes or beards, to create an output that corresponds to the original input.[15] Multiple iterations of this process enhance the model's predictive ability. This process is coupled with linking the images with text descriptions of the image, so that the generative model can output an image that matches the textual description in the output. Variational encoders differ from other models, in that they use variational inference to predict the qualities of intractable attributes of an image using statistical probability.[16]

A GAN is a model which also has two parts, but as the name suggests, these two parts are adversarial neural networks. These two networks are trained simultaneously and are made to compete with each other.[17] The first network is given a random input and is trained via backpropagation. Backpropagation is a process where, after an input is given and an output is generated, the error between the actual output and the desired output is sought to be corrected by re-feeding the actual output in the network and adjusting weights.[18] The second network is discriminative, and is tasked with distinguishing the real image from the generated artificial

---

[11] Osval Antonio Montesinos López, Abelardo Montesinos López and José Crossa, *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (Springer 2022) 399.

[12] Ajay Bandi and others, 'The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges' (2023) 15(8) Future Internet 260 <www.doi.org/10.3390/fi15080260> accessed 8 July 2025.

[13] Jessica L Gillotte, 'Copyright Infringement in AI-Generated Artworks' (2019) 53 UC Davis L Rev 2663 <www.lawreview.law.ucdavis.edu/sites/g/files/dgvnsk15026/files/media/documents/53-5_Gillotte.pdf> accessed 8 July 2025.

[14] Jaan Altosaar, 'Tutorial- What Is a Variational Autoencoder?' (*Jaan Li*, 16 August 2016) <www.jaan.io/what-is-variational-autoencoder-vae-tutorial/> accessed 15 February 2024.

[15] Miguel Mendez, 'The Theory behind Variational Autoencoders' (*Miguel Mendez*, 19 January 2019) <www.mmeendez8.github.io/2019/01/19/vae-theory.html> accessed 6 February 2024.

[16] 'Variational AutoEncoders' (*GeeksforGeeks*, 20 July 2020) <www.geeksforgeeks.org/variational-autoencoders/> accessed 6 February 2024.

[17] Ian J Goodfellow and others, 'Generative Adversarial Networks' (*Cornell University*, 10 June 2014) <www.arxiv.org/abs/1406.2661> accessed 7 September 2023.

[18] Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning* (MIT Press 2016) 200–202.

image. The process of training goes on until the discriminative model cannot distinguish between the real and the artificial.[19]

This preliminary understanding is relevant to our present discussion, since the manner of training and its precise modalities will help — (1) determine whether there was any infringement due to storage and reproduction, (2) in understanding the perspectives to comprehend generative models, and (3) in determining recommendations for a regulatory framework.

## RECENT LAWSUITS CONCERNING GENERATIVE AI

### a. GETTY IMAGES v. STABILITY INC.

In early 2023, Getty filed a lawsuit against Stability AI (hereinafter 'Stability'), the developer of the popular image generation tool Stable Diffusion.[20] Getty Images (hereinafter 'Getty'), a popular photo and media company, which conducts its business in aggregating photographs and licensing them as stock photos.[21] A large chunk of Getty's business is entering into licensing agreements, with people willing to use photos that are uploaded on their website by photographers, and thus, Getty also acts as a platform intermediating the licensing process.[22] Getty Images, also registers each of its items for copyright protection. A unique feature of Getty Images is that each of their photograph has a watermark on the centre-left of the image of their logo and the name of the photographer. Moreover, the images also have an alt-text, which describes the contents of the image.[23]

The gravamen of the complaint lies in Stability's unauthorised use of copyright-registered images in training their generative model. Getty Images alleges that there was a violation of copyright in the images and the text captions accompanying the images which were contended to be original works protectable by US Copyright Law.[24] Getty Images claims that their service provides high-quality images on a variety of subjects, which makes them more desirable to use as training data,[25] since low-quality images do not provide enough fidelity to allow the generative models to collect all necessary data-points.[26] Moreover, Getty Images always

[19] ibid.
[20] *Getty Images (US), Inc v Stability AI, Inc* No 1:23-cv-00135-UNA.
[21] ibid 18.
[22] ibid 19, 20.
[23] ibid 25, 31.
[24] ibid 23, 24.
[25] ibid 18.
[26] Alberto Rizzoli, 'An Introductory Guide To Quality Training Data For Machine Learning' (*V7 Labs*, 11 July 2022) <www.v7labs.com/blog/quality-training-data-for-machine-learning-guide> accessed 17 February 2024.

captions their photographs with highly descriptive text covering all significant parameters of an image. A sample caption is "*A section of Lake Oroville is seen nearly dry on August 19, 2014 in Oroville, California. As the severe drought in California continues for a third straight year, water levels in the State's lakes and reservoirs is reaching historic lows. Lake Oroville is currently at 32 percent of its total 3,537,577-acre feet.*"[27] Captions like these, combined with a high-resolution image, are rich resources for training.

Stable Diffusion, uses a combination of Variational Autoencoder, a text-image encoder, and a modified version of a Convolutional Neural Network.[28] The training process of Stable Diffusion allegedly involves copying and storing of the images along with the text descriptions of those images. The images then go through a process of forward feeding and backpropagation, where images are blurred by the addition of noise, and the model is tasked with denoising the image in order to create an output that matches the description.[29] The exact mechanism of "learning" occurs using noising and de-noising the inputs, and changing the values of weights to reduce the bias in the model, as explained above in Part 2. The dataset of the works used to train the model was provided by a German organisation LAION, which crawled the web and collated links to photographs along with the text description available.[30] The database exists in a tabular form where the language, text description, metadata, and the URL to the image is collated.[31] The database, also allegedly contained the registered works exclusively licensed to Getty.

The main copyright claim in this case is unauthorised reproduction, and creation of derivative works of an infringing nature. Getty has also accused Stability of tampering with the rights management information on the photographs exclusively licensed to Getty, with there being many cases of outputs that clearly show the "Getty Images" logo watermarked onto outputs

---

[27] Justin Suvilian, 'Statewide Drought Takes Toll On California's Lake Oroville Water Level' (*Getty Images*, 19 August 2014) <www.gettyimages.in/detail/news-photo/section-of-lake-oroville-is-seen-nearly-dry-on-august-19-news-photo/453834006> accessed 17 February 2024.

[28] Robin Rombach and others, 'High-Resolution Image Synthesis with Latent Diffusion Models' (*Cornell University,* 13 April 2022) <www.arxiv.org/abs/2112.10752> accessed 6 February 2024); Jay Alammar, 'The Illustrated Stable Diffusion' (*Jalammar,* 4 October 2022) <www.jalammar.github.io/illustrated-stable-diffusion/> accessed 6 February 2024.

[29] Jay Alammar, 'The Illustrated Stable Diffusion' (*Jalammar,* 4 October 2022)<www.jalammar.github.io/illustrated-stable-diffusion/> accessed 6 February 2024.

[30] *Getty Images (US), Inc v Stability AI, Inc* No 1:23-cv-00135-UNA, 38; Andy Baio, 'Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator' (*Waxy*, 30 August 2022)<www.waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/> accessed 17 February 2024; 'Large Scale Artificial Intelligence Open Network' (*LAION*)<www.laion.ai/> accessed 6 September 2023.

[31] 'Datasets at Hugging Face' (*Hugging Face*) <www.huggingface.co/datasets/laion/laion2B-multi-aesthetic/viewer/default/train> accessed 6 September 2023.

generated by Stable Diffusion, with one being substantially similar to an original image in which Getty owns copyright.


*Figure 1. Original Image*

*Figure 2. Output from Stable Diffusion*

### b. ANDERSEN ET AL v. STABILITY INC.

This is another lawsuit filed by three different artists, where all three of them allege copyright infringement in their artistic compositions by Stability in training Stable Diffusion.[32] The main claim in this lawsuit pertains to unlawful distribution, creation of unauthorised derivative works, unlawful reproduction of the works, and performance of the work in public. On 30 October 2023, the Court rejected all other claims involved in the suit, except the claim for copyright infringement against Stability Inc. The Court, however, granted the plaintiffs leave to amend the plaint to pinpoint with higher specificity how the copyrighted images were used. The Court admits that the theory of liability is an evolving one, since the plaintiffs are yet to examine third parties who can assist in the actual manner of use of the images. Thus, the reason for the current status of this trial seems to be due to the lack of specificity in the plaintiffs' claims and the fact that various third parties who have material information are yet to be served subpoenas, rather than on the merits of the case.

[32] *Andersen v Stability AI Ltd* No 3:23-cv-00201.

### c. KADREY v. META PLATFORMS

This case concerns copyright infringement claims in training data used in Meta's LLaMA, a large language model. Although, this case does not relate to visual generative models, it is notable in that it engages in the treatment of the copyright status of training data. Three authors who had works registered with the copyright registry of the United States, filed this plaint, alleging infringement of their works. Meta used the Books3 database for training its model, and it is admitted by the makers of the database, that it scraped books from "shadow libraries", like Bibliotek, which host pirated books and infringing material.[33] The plaintiffs claimed infringement of exclusive rights of reproduction, making unauthorised copies, communication to public, and contravention of copyright management information.

The Court, in a motion to be dismissed, largely decided in Meta's favour.[34] The claim against Meta was that their use of the plaintiff's works was an infringement of the plaintiff's copyright, since such use violated their exclusive right to create a derivative work.[35] This court further held that the plaintiffs would have to prove that the outputs generated by LLaMa were substantially similar to the inputs. The judgment was thus decided on the claim that the creation of the generative model in itself was a derivative work. This case overlooks the fact that Meta used copyrighted material and possibly stored it in a material form. American copyright law provides an exclusive right of reproduction which vests with authors.[36] This includes the right to make copies, which are "*material objects, other than phonorecords, in which a work is fixed…*".[37] Fixation has been referred to mean stable or permanent, such that it can be perceived, communicated or reproduced, for a duration which is not transitory.[38] Therefore, unauthorised storage of copyrighted works in digital form must be understood as infringement. Courts have held that unauthorised storage of copyrighted works in digital form can amount to infringement.[39] The pertinent question that should have been asked is whether there was actual

---

[33] Moohita Kaur Garg, 'Meta flouted copyrights to train its AI 'Llama' despite warning from lawyers, claims lawsuit by authors' (*WION*, 13 December 2023) <www.wionews.com/technology/meta-flouted-copyrights-to-train-its-ai-llama-despite-warning-from-lawyers-claims-lawsuit-669148> accessed 17 February 2024); Sarah Brady, 'Authors Accuse Meta of 'knowingly' Training AI with Copyrighted Books' (*Verdict,* 13 December 2023) <www.verdict.co.uk/authors-accuse-meta-of-knowingly-training-ai-with-copyrighted-books/> accessed 17 February 2024.

[34] *Kadrey v Meta Platforms, Inc* 2023 US Dist LEXIS 207683, 2023 WL 8039640.

[35] *Kadrey v Meta Platforms, Inc* No 3:23-cv-03417.

[36] 17 USC §106(1) (2024).

[37] 17 USC §101 (2024).

[38] Melville B Nimmer, *Nimmer on Copyright: A Treatise on the Law of Literary, Musical and Artistic Property, and the Protection of Ideas* (M Bender 1965) 8.02.

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

material storage by Meta to help prove infringement, if any, similar to the proceedings in Getty Images (supra).

### d. CONCORD MUSIC GROUP, INC. v. ANTHROPIC PBC[40]

This case concerns infringement claims made against Anthropic, an AI-research company formed by a break-away group of ex-employees at OpenAI, which aims at creating a generative model with a focus on AI safety.[41] Anthropic developed its own large language model, named Claude, which functions in a manner similar to other text generating models. Although, this dispute does not focus on visual generative models, it provides an example of how outputs of generative models can be substantially similar to copyrighted works. The complaint showcases the manner of substantially similar reproductions, that a generative model can generate when prompted. For example, the plaintiffs prompted Claude to "*Write … a song about the death of Buddy Holly*", and Claude responded by generating the lyrics to the song "American Pie" by Don McLean, the copyright in which vests with the plaintiff. The answer generated by Claude shows the verses in the said song, but their order is rearranged, while still retaining the exact sequence of words within a verse. There are minor, inconsequential rearrangements to the sentence structures, but the answer is substantially similar to the actual lyrics.[42] Another prompt by the plaintiffs was "*Write a song about moving from Philadelphia to Bel Air*" which is admittedly a generic statement, with no identification of authors, or the names of songs or characters, and yet again, Claude generates the exact lyrics to the song "The Fresh Prince of Bel-Air" authored by Will Smith.[43] Thus, we see that one of the major challenges that copyright infringement claimants face is that of informational asymmetry, since only the developers of the generative models know the datasets used and the precise method and manner of training.[44] Courts must strive to seek the contents of the datasets used and the manner of their use to effectively deal with copyright infringement claims.

---

[40] *Concord Music Group, Inc v Anthropic PBC* No 3:23-cv-01092.
[41] James Vincent, 'Google Invested $300 Million in AI Firm Founded by Former OpenAI Researchers' (*The Verge*, 3 February 2023) <www.theverge.com/2023/2/3/23584540/google-anthropic-investment-300-million-openai-chatgpt-rival-claude> accessed 24 February 2024.
[42] *Concord Music Group, Inc v Anthropic PBC* No 3:23-cv-01092, 73.
[43] ibid.
[44] Kyle Barr, 'GPT-4 Is a Giant Black Box and Its Training Data Remains a Mystery' (*Gizmodo*, 16 March 2023) <www.gizmodo.com/chatbot-gpt4-open-ai-ai-bing-microsoft-1850229989> accessed 17 February 2024.

# PERSPECTIVES ON TRAINING DATA AND COPYRIGHT INFRINGEMENT

## I. TRAINING DATA AS COLLAGE

A parallel may be drawn between collage creation, and training a generative model to create a new visual work.[45] It is tempting to draw an analogy between a collage and a generative model; after all, a generative model collates a large database of images and generates fresh works on the basis of the data it receives. A collage, which complies with copyright law, must be understood as a collection of multiple works, requiring a particular creative vision, which upon execution becomes a collation of carefully selected works which are original enough to convey a new idea. It may be argued that a visual generative model also works similarly. It ingests multiple images, selects the characteristics that are required by the input, and creates a new image on the basis of the many pictures it originally ingested in order to output an image with a new idea.

However, such a conception is not accurate. A collage consists of pieces of works, which are collated and made into an integrated whole, which then creates a novel whole. Although the analogy is tempting in that a generative model is trained on various images and it creates a new image using those, this logic misses the nuances of the process of training and the manner in which neural networks function. A more valid analogy would be with papier-mâché, where multiple layers of paper clump together to create a new work, and the identity of the individual bits of paper is lost in service of the resultant whole.[46] While training, the model is taught to recognise various attributes of an image. For example, LLaMa, a textual generative model, boasts as many as 65 million parameters, which are characteristics of a possible image, like face shape, size or colour.[47] Resultantly, a generative model "learns" using feedforwarding and backpropagation, to first derive latent characteristics of the input images it consumes, and stores the data it learns from the text-image connection. Furthermore, generative models use noising and then denoising images, to allow the model to identify characteristics better by way of abstraction of those characteristics. In summation, a generative model cannot be equated

---

[45] Nettrice Gaskins, 'Romare Bearden, Mechanical Reproduction & Generative AI' (*Medium*, 1 September 2023) <www.nettricegaskins.medium.com/romare-bearden-mechanical-reproduction-generative-ai-6df2c4f2750a> accessed 6 September 2023.

[46] Matthew Sag, 'Copyright Safety for Generative AI' (2023) 61(2) Hous L Rev 321 <www.papers.ssrn.com/sol3/papers.cfm?abstract_id=4438593> accessed 10 July 2025.

[47] 'Introducing LLaMA: A foundational, 65-billion-parameter language model' (*Meta,* 24 February 2023) <www.ai.meta.com/blog/large-language-model-llama-meta-ai/> accessed 6 September 2023; Jason Brownlee, 'What is the Difference Between a Parameter and a Hyperparameter?' (*MachineLearningMastery*, 17 June 2019) <www.machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/> accessed 6 September 2023.

with a collage machine, since it identifies characteristics of images to create new imagery based on the understanding of characteristics of various kinds of images and the interface of these characteristics. Moreover, generative models go through a process of gaining "experience", which provides them with a degree of autonomy in the exact manner of representation of various attributes of the output image. Thus, any judicial treatment of collages and their interface with copyright ought not to apply to cases claiming copyright infringement in the use of training data.

## II. DATA MINING, FAIR USE AND GENERATIVE MODELS

Data mining has received considerable judicial treatment and has been held to qualify as fair use. Data mining broadly refers to the process of improving future decisions, by collecting data from past events and finding patterns in that data.[48] For example, US law, in Authors Guild v. Google[49] and HathiTrust v. Google[50] (hereinafter 'Authors Guild cases') upheld the practice of textual data mining for the purposes of research as fair use.

The American fair use standard is a four-pronged test which involves an inquiry into – *"(a) the purpose and character of the use, (b) nature of the copyrighted work; (c) amount and substantiality of the portion used and (d) effect of the use upon the potential market or value of the copyrighted work*." (hereinafter 'Four-Factor Test').[51] In these cases, Google and Hathi Trust, two digital libraries gained access to complete versions of copyrighted books, scanned them, and provided a full text search for the books in question. The issue of the use of copyrighted books to extract metadata was objected to by authors who held copyright in those books. The primary factor that led to a finding in favour of fair use was the academic and research value of the use which made it transformative.[52] The Court held that this use by Google increased the ease with which relevant materials could be sourced, thereby, aiding research and writing. This, in conjugation with the constitutional purpose to promote arts and sciences by way of copyright, made the use transformative.[53] Further, relying on its decisional history, the Court observed that when a finding of transformative use exists, it outweighs any market harm

[48] Tom M Mitchell, 'Machine Learning and Data Mining' (1999) 42(1) COMMUN. ACM 32 <www.cs.cmu.edu/~tom/pubs/cacm99_final.pdf> accessed 10 July 2025.
[49] *Authors Guild v Google, Inc* 804 F 3d 202, 2015 US App LEXIS 17988.
[50] *Authors Guild v HathiTrust* 755 F 3d 87.
[51] 17 USC §107 (2024).
[52] *Authors Guild v Google, Inc* 804 F 3d 202, 2015 US App LEXIS 17988, 216.
[53] ibid 217.

that may be caused to the plaintiffs.[54] The Court, thus, also found that the search function created by Google and Hathitrust would not diminish the market for the plaintiff's books, since the search function did not make the entirety of the book free to read. The Court held that users of Google's and HathiTrust's products would have to purchase the books in order to view their complete and contextualised contents.[55] The Court held in these cases that the nature and character of the use was to promote research, and development of humanities research. It also emphasised the impact of the use on scholarly research, promotion of the arts, science, and education.

The same justifications cannot be invoked for visual generative models. Firstly, the reasoning that favours a fair use finding is absent in the case of generative models, and secondly, the Authors Guild cases do not serve as viable precedents. There has been a great deal of opinion on the benefits of visual generative models, however, in the ultimate analysis, one ought to balance the negative with the positive. The most common reasons cited are that visual generative models will help augment artistic creativity, enhance access to art for commercial purposes, or that it would "democratise access to art".[56] These reasons do not cause any social benefit, such that humankind might progress further with, and neither does the use of generative models promote creativity. One might argue that their use would create a substitute to creativity by supplanting the incentive to protect human creations with readily available machine-generated art which substitutes human-made works.[57] Furthermore, the use of training data would harm human artists, whose works diminish in value as a result of the creation of these models, which ought to be a relevant balancing consideration in holding against fair use.[58] The factual matrix in the Authors Guild cases, was such that no substitute to the market of the plaintiffs existed such that there would be a loss to the market of the plaintiff's works, and this was an important consideration in the holding finding fair use. The same does not hold true for the use of training data in generative models, as seen in Getty Images v. Stability AI, where the images in the LAION database contained copyright-protected images, thus, creating losses for Getty.[59] The United States Supreme Court has observed that the market harm factor is

---

[54] ibid 219; *Cariou v Prince* 714 F 3d 694 (2d Cir 2013).
[55] *Authors Guild v Google, Inc* 804 F 3d 202, 2015 US App LEXIS 17988, 224.
[56] 'Exploring the Benefits of Generative AI' (*Talespin*, 31 July 2023) <www.talespin.com/reading/exploring-the-benefits-of-generative-ai> accessed 7 September 2023.
[57] Nelson Granados, 'Human Borgs: How Artificial Intelligence Can Kill Creativity And Make Us Dumber' (*Forbes*, 31 January 2022) <www.forbes.com/sites/nelsongranados/2022/01/31/human-borgs-how-artificial-intelligence-can-kill-creativity-and-make-us-dumber/> accessed 18 February 2024.
[58] Pierre N Leval, 'Toward a Fair Use Standard' (1990) 103 Harv L Rev 1105 <www.tandfonline.com/doi/full/10.1080/10811680.2020.1767419> accessed 10 July 2025.
[59] *Getty Images (US), Inc v Stability AI, Inc* No 1:23-cv-00135-UNA.

"*undoubtedly the single most important element of fair use*". A finding of transformative use may, however, result in an outcome in favour of fair use.[60]

Therefore, it may be argued that none of these apparent benefits plausibly outweigh the harm caused to artists whose works are exploited and their market being replaced,[61] nor does it further any progress to the arts and science insofar as human artists are harmed, and as such the Authors Guild cases ought not be used as precedent.[62] Further, such use goes against the raison d'être of copyright law, which is to promote creative work by incentivising the creation of works which are artistic, or beneficial to humanity.[63] Furthermore, it must be observed that the Authors Guild cases grant a data mining exception for use where no original expression is created. The creation of a search function which aids in searching for relevant books is a functional, non-expressive use, and as such the Authors Guild cases ought to be distinguished and held inapplicable to generative models, which produce expressive results. Any application of the Authors Guild precedents would miss out on the nuances of the training and functioning of generative models.[64] A balance between the rights of the copyright holders, and the need to promote technological advancement must be struck. Adding to the chaos, is the unpredictable nature of fair use inquiries, both in the United States and in India.[65] American law lays down the four-factor test, and each factor is weighed holistically, regard being to the attendant circumstances, where each factor does not hold equal weight and is not always equally crucial to the fair use analysis.[66] American courts have gone so far as to state that the fair use doctrine is "*the most troublesome [doctrine] in the whole law of copyright*", calling it a "billowing white goo."[67] Indian law, on the other hand, refers to Section 52, where the precise import of exceptions is stated, thus, creating a slightly more precise framework. Although Indian case law refers to the American standard, the statute's bare language takes precedence given that Section 16 mandates copyright emanation only from the provisions of the statute, and from no

---

[60] *Harper & Row Publishers v Nation Enter* 471 US 566 (1985).
[61] *Campbell v Acuff-Rose Music, Inc* 510 US 569 (1994).
[62] *Google LLC v Oracle Am, Inc* 141 S Ct 1183.
[63] Wenhong Qu, 'The Humanistic Value of Knowledge Economy and Law' (2021) 4(5) Proceedings of Business and Economic Studies 105 <www.ojs.bbwpublisher.com/index.php/PBES/article/view/2668> accessed 10 July 2025.
[64] David W Opderbeck, 'Copyright in AI Training Data: A Human-Centered Approach' (2023) 76(4) Okla L Rev 976–981 <www.digitalcommons.law.ou.edu/cgi/viewcontent.cgi?article=2305&context=olr> accessed 10 July 2025.
[65] Jenny Quang, 'Does AI Training Violate Copyright Law?' (2021) 36 Berkeley Tech LJ 1422 <www.btlj.org/wp-content/uploads/2023/02/0003-36-4Quang.pdf> accessed 10 July 2025.
[66] 17 USC §107; *Google LLC v Oracle Am, Inc* 141 S Ct 1183, 209 L Ed 2d 311, 2021 US LEXIS 1864, 2021 USPQ2D (BNA) 391, 28 Fla L Weekly Fed S 727, 2021 WL 1240906.
[67] *VHT, Inc v Zillow Grp, Inc* 918 F 3d 723, 739 (9th Cir 2019).

other sources.[68] However, Indian courts frequently refer to the American standard as persuasive precedents.[69] Thus, it is recommended that legislative intervention carve out a framework to allow training of generative models while allowing copyright owners their rights.

### III. GENERATIVE MODELS AND THEIR BIOMIMETIC CHARACTER

It is also tempting to draw a parallel between generative models and human creativity. After all, humans also sample a large number of art and then learn from their characteristics, which knowledge is synthesised and then used to create an original work, which is capable of being granted copyright protection, and which is also not an infringement of any rights in the earlier artworks which serve as inspiration. However, such a paradigm is too simplistic.[70] Copyright law prevents material reproductions and copying. In both American law and Indian law, making of copies in a computer or other device which allows for storage, amounts to infringement. The difference between human learning and the use of works for training generative models lies in this fixation in a material form, which constitutes an exclusive right capable of infringement.

## PROGNOSIS FOR INDIA

The recent lawsuits in various jurisdictions about generative artificial intelligence share a common thread. The most potent copyright infringement claims that can be made with respect to use of training data and creation of datasets facilitating such training are in (i) reproduction or making copies, (ii) communication to public, and (iii) circumvention of copyright management information.

### RIGHT TO REPRODUCTION

The right to reproduction in India is provided for by Section 14 of the Copyright Act, 1957 (hereinafter 'Act, 1957').[71] More pertinently, Section 14(c)(i)(A) deals with artistic works and creates an exclusive right to the owner to store the work in "*any medium by electronic or other means*", which creates two shades of analysis. The first shade pertains to the claim that the output made by a generative visual model is a reproduction of the training data, and the second

[68] *Tips Industries Ltd v Wynk Music Ltd* 2019 SCC OnLine Bom 13087.
[69] *Masters & Scholars of University of Oxford v Rameshwari Photocopy Services* 2016 SCC OnLine Del 6229.
[70] Jenny Quang, 'Does AI Training Violate Copyright Law?' (2021) 36 Berkeley Tech LJ 1414 <www.btlj.org/wp-content/uploads/2023/02/0003-36-4Quang.pdf> accessed 10 July 2025.
[71] Copyright Act 1957, s 14(c)(i)(A).

shade is the claim against unauthorised storage. The former becomes pertinent when a generative model is asked to generate an image of a copyrighted work, and it does so faithfully (such as the reproduction of a cartoon character, see Figure 3 below), or where it creates an output which is substantially similar to an earlier work. The latter becomes relevant where it is proved that while training the generative model, the developers stored copies of the copyrighted works. In cases of reproduction claims in outputs of the generative model, current law follows the principle of substantial similarity.[72] The two works must be compared and substantial similarity must be established. If, *prima facie,* the works appear to be near-similar to each other, the initial and rebuttable burden of proof of reproduction is cast on the defendant, which can be resiled against. For a claim against violation of reproduction rights to succeed, it must be proved that there is an objective similarity between the earlier work and the allegedly infringing work. Where such an objective similarity is found, this would be evidence of a causal connection.[73] Similarity is also determined by the impression that the comparison yields in a layman's eyes.[74] In some cases, like in V. Govindan v. E. M. Gopalakrishnan, the Court also places the onus of proving that there was no copying on the defendant by proving that the work done was made originally by starting from scratch.[75] Reproduction may also be proved by comparing the overall artistic feel of the two works. This was exemplified in Thomas v. Bradbury, Agnew & Co., Ltd., where a visual character was adapted into a 3-dimensional form by way of costume, the Court opined that the transformation replicated the feel of the character.[76]

In this respect, generative models generally do not constitute reproduction. In an empirical study conducted to check the similarity between training data images and the outputs in generative models, it was found that only 109 of 175 million outputs would constitute substantial similarity, which is also a subjective analysis.[77] Therefore, infringement potential would be quite low. However, there are "edge cases", which allow a possibility for infringement by reproduction/copying to take place. For example, artists protested the AI art movement by asking Dall-E and MidJourney to create images of Mickey Mouse.[78] Although

---

[72] *RG Anand v Delux Films* (1978) 4 SCC 118 [45], [46].
[73] *Francis Day & Hunter Ltd v Bron* [1963] 2 WLR 868.
[74] *Ram Sampath v Rajesh Roshan* 2008 SCC OnLine Bom 370.
[75] *V Govindan v EM Gopalakrishna Kone* 1954 SCC OnLine Mad 368.
[76] *Thomas v Bradbury, Agnew & Co Ltd* 2 K B 627 (1906).
[77] Nicholas Carlini and others, 'Extracting Training Data from Diffusion Models' (*Cornell University*, 30 January 2023) <www.arxiv.org/abs/2301.13188> accessed 7 September 2023.
[78] Mikael Thalen, 'Artists Fed up with AI-Image Generators Use Mickey Mouse to Goad Copyright Lawsuits' (*The Daily Dot*, 19 December 2022) <www.dailydot.com/debug/ai-art-protest-disney-characters-mickey-mouse/> accessed 7 September 2023.

no copyright currently subsists in Mickey Mouse,[79] it would constitute infringement in cases where a copyrighted work is reproduced and would be squarely covered by the holding in Thomas v. Bradbury, Agnew & Co., Ltd. and R.G. Anand v. Delux Films, since there is substantial similarity, although, there is a change in the medium of the work.

Furthermore, a case in point for infringement would be the substantial similarity in the copyrighted photo belonging to Getty and the output generated by Stable Diffusion in Getty Images v. Stable Diffusion (supra). A fair use finding in the United States would also have to overcome the holding in Andy Warhol Foundation for the Visual Arts, Inc. v. Lynn Goldsmith, et al.,[80] where a claim that the works created by Andy Warhol were transformative was rejected since there was substantial similarity between the prior work and Warhol's work,[81] given that there is at least one instance where the output by generated by Stable Diffusion is substantially similar to an identifiable input.

*Figure 3. Mickey Mouse made using Stable Diffusion*[82]

In the case of storage of works, the case of Tips Industries Ltd. v. Wynk Music Ltd. is instructive. The case held that the storage of sound recordings by the defendants was infringement of the exclusive right to storage under Section 14(e)(i). Furthermore, the case also established that fair use under the Indian copyright law, is restricted to the statute's language.[83] Therefore, the crucial question that remains to be seen is whether the training process of a

[79] Anna Gordon, 'Mickey Mouse is Now in the Public Domain' (*Time*, 2 January 2024) <www.time.com/6551496/mickey-mouse-public-domain-steamboat-willie/> accessed 18 February 2024.
[80] *Andy Warhol Found for the Visual Arts, Inc v Goldsmith* 143 S Ct 1258, 215 L Ed 2d 473, 2023 US LEXIS 2061, 598 US 508, 2023 USPQ2D (BNA) 603, 29 Fla L Weekly Fed S 805 (US May 18, 2023).
[81] *Andy Warhol Found for the Visual Arts, Inc v Goldsmith* 143 S Ct 1258, 215 L Ed 2d 473, 2023 US LEXIS 2061, 598 US 508, 2023 USPQ2D (BNA) 603, 29 Fla L Weekly Fed S 805 (US May 18, 2023).
[82] Sakipooh, 'Mickey' (*Reddit*, 2022) <www.reddit.com/r/StableDiffusion/comments/xtg023/mickey/> accessed 7 September 2023.
[83] *Tips Industries Ltd v Wynk Music Ltd* 2019 SCC OnLine Bom 13087.

generative model requires downloading and storage of data in order to establish a violation of rights in the Indian context. This can prove to be a herculean task given that the exact methods and processes of training may not have been disclosed.[84] However, it prima facie seems probable that storage was involved, given the process of training, given that a neural network performs various functions, such as noising and de-noising on the input training images, as explained hereinabove in Part 2.

## COMMUNICATION TO PUBLIC

Training datasets are seldom created by the creators of generative models themselves. As seen in the case of Getty Images v. Stability AI, a German association called LAION, made a collation of multiple links along with their textual description of the images. There is quite a lot of judicial opinion about linking imagery amounting to communication to public. For example, under EU law, the decision in GS Media v. Sanoma Media[85] currently holds the field. The holding states that hyperlinking is not *per se* communication to public, but if the person making that hyperlink knows about that the link hosts infringing work, it would amount to communication to public.[86] This has been criticised as putting an unnecessary and impractical burden on the person hyperlinking the content.[87] American law on this point has been largely settled post the rulings in Kelly v. Arriba Soft Corp (hereinafter 'Kelly')[88], and Perfect 10, Inc. v. Amazon Inc. (hereinafter 'Perfect 10').[89] In Kelly, the use of deeplinking (a form of hyperlinking) of images by a search engine was held to be sufficiently transformative, since it created a novel method to search images. It was held to be fair use also because of negligible market harm to the other party. In Perfect 10, the Court, being seized of a matter of displaying an image as a thumbnail, held that it was fair use on grounds of transformative use, sufficient public benefit, and also due to the fact that the thumbnails were compressed images hosted on the web using HTML copies. Therefore, a fair use analysis of hyperlinks is resorted to under American law.

German law, on the other hand, considers hyperlinking to be an infringement. In Decision I-20 U 42/11 Dusseldorf Court of Appeal 8 October 2011, the Court held that hyperlinking could

---

[84] Kyle Barr, 'GPT-4 Is a Giant Black Box and Its Training Data Remains a Mystery' (*Gizmodo*, 16 March 2023) <www.gizmodo.com/chatbot-gpt4-open-ai-ai-bing-microsoft-1850229989> accessed 17 February 2024.
[85] *GS Media BV v Sanoma Media Netherlands BV and others* [2016] Bus LR 1231.
[86] ibid 49, 51.
[87] Tobias Cohen Jehoram, 'European Court of Justice: hyperlinks to unauthorized content may infringe copyright' (*WIPO,* 16 December 2016) <www.wipo.int/wipo_magazine/en/2016/06/article_0007.html> accessed 7 September 2023.
[88] *Kelly v Arriba Soft Corp* 336 F 3d 811 (9th Cir 2003).
[89] *Perfect 10, Inc v Amazon.com, Inc* 508 F 3d 1146 (9th Cir 2006).

result in copyright infringement if no permission was taken. Belgian case law, in Copiepresse v. Google Inc.[90] also considers hyperlinking copyright infringement.[91] Indian courts have not discussed this issue; however, some cursory treatment is found in Blueberry Books v. Google India (P) Ltd.,[92] which concerned a hyperlink which would give access to a copyright-protected image on Amazon, was available for download. However, the Delhi High Court did not lay down the law of hyperlinking and copyright infringement, since it concluded that the download feature was available only in the United States, and hence, refused to entertain the plaint. An interlocutory appeal was preferred with a division bench which passed its judgment in 2016 and reversed the single judge's decision and allowed the petition to the extent that some parties were re-arraigned in the suit. However, no proper discussion of the merits was made.[93] Therefore, the issue of communication to public is dealt very differently in various jurisdictions and is therefore, a possible hurdle for generative model creators.

## VIOLATION OF RIGHTS MANAGEMENT INFORMATION

The most blatant violation of the distortion of rights management information can be seen in Getty Images, where the watermark has been reproduced in a manner sufficiently similar to the original. The watermark consists of the logo of Getty and the name of the photographer. Indian law, on this point is provided for in Section 65B. Section 2(xaa) of the Act, defines the term as including title, name of author, metadata, etc. This can be a potential claim against the developers of generative models as well as the creators of databases.

## SECTION 52: FAIR DEALING

Section 52 of the Act, 1957 provides for a list of uses which do not amount to infringement. The language of the section uses "namely", indicating that these are largely exhaustive.[94] However, through judicial pronouncements, the scope of the section has been somewhat enlarged. An example of this would be the adoption of the concept of transformative use which does not appear in the statute's language.[95] That being said, any claim of fair dealing must

---

[90] *Copiepresse v Google Inc* [2007] 23 CLSR 82-85, [2007] 23 CLSR 290-293, High Court of Brussels.
[91] Philippe Laurent, 'Copiepresse SCRL & Alii v. Google Inc. – In Its Decision of 5 May 2011, the Brussels Court of Appeal Confirms the Prohibitory Injunction Order Banning Google News and Google's "in Cache" Function' (2011) 27(5) Computer Law & Security Review 542 <www.sciencedirect.com/science/article/abs/pii/S0267364911001208> accessed 10 July 2025.
[92] *Blueberry Books v Google India Pvt Ltd* 2013 SCC OnLine Del 4805.
[93] *Blueberry Books v Google India Pvt Ltd* 2016 SCC OnLine Del 3338.
[94] Copyright Act 1957, s 52(1).
[95] *RMC Project Management International v Whizlabs Software (P) Ltd* 2023 SCC OnLine Del 5169.

make reference to the statutory parameters and must satisfy the pertinent requirements of any provision under which a finding for fair dealing is claimed.

Under current law, there are few possible applicable cases where a relevant fair use defence can be claimed. The first of these is Section 52(1)(a), which protects fair dealing in works not being computer programmes, where the use is for private or personal research.[96] This would be inapplicable to the current case, since the use is not private or personal, but commercial in nature, since many generative models have paid versions available.[97]

Secondly, as far as communication to public by makers of datasets, such as LAION, is concerned, Section 52(1)(c) might be invoked. Section 52(1)(c) provides that a "transient and incidental storage" of a work, where the purpose of use is providing links or access to that work is not infringement, provided the copyright owner has not explicitly prohibited such use. Moreover, where the copyright owner has served a notice asking for such use to be ceased, or if a court has ordered prohibition on such use, the person storing the copyrighted work must comply and cease from storing the work.[98] However, this is a difficult defence to mount, since many sources from which the makers of training datasets have obtained images from, have usage policies or terms of service in place which prohibit unauthorised copying of the works. Getty Images' licensing service would be a good case in point.[99] Moreover, it must be proved that the storage was merely transient and incidental to the purpose of providing access. The "transient and incidental storage" also appears in Section 52(1)(b), wherein, any such storage, which is purely a result of a technical process of communication to public can be invoked as a defence, if the manner of communication to public of the dataset meets the abovementioned requirement.[100]

No judicial treatment of the meaning of "transient and storage" exists in India currently. However, as per the Merriam-Webster Dictionary, the word transient means "*passing especially quickly into and out of existence*", and as such, the storage must be for an ephemeral duration.[101] Reference may be made to Cartoon Network LP, LLLP v. CSC Holdings, Inc., wherein the issue of whether storage of data for a duration of 1.2 seconds was a "copy" was

---

[96] Copyright Act 1957, s 52(1)(a)(i).

[97] Emilia David, 'Stability AI Announces Paid Membership for Commercial Use of Its Models' (*The Verge*, 20 December 2023) <www.theverge.com/2023/12/19/24008149/stability-ai-paid-subscription-commercial-rights-safety> accessed 18 February 2024.

[98] Copyright Act 1957, s 52(1)(c).

[99] 'Getty Images Site Terms Of Use' (*Getty Images*, February 2024) <www.gettyimages.in/company/terms> accessed 18 February 2024; 'Terms of Service' (*Imgur*) <www.imgur.com/tos> accessed 18 February 2024.

[100] Copyright Act 1957, s 52(1)(b).

[101] 'Transient' (*Merriam-Webster Dictionary*) <www.merriam-webster.com/dictionary/transient> accessed 18 February 2024.

answered in the negative, since it was "fleeting" storage.[102] It is submitted that fleeting, being a synonym of transient, must be given the same interpretation.

Fair dealing found judicial treatment in Super Cassettes Industries Ltd. v. Hamar Television Network (P) Ltd.,[103] where the Delhi High Court, while seized of a matter in which the defendants invoked the defence of Section 52(1)(a) of the Copyright Act, 1957, observed that the question of fair dealing is a fact-based one, and no rigid standard regarding a fair dealing finding can be crafted. The Court determined a test similar to the American four-factor test. The Court recommended that the quantum of the copying, both qualitative and quantitative, must be accounted for.[104] The Court adopted a standard similar to the "heart of the work" standard,[105] where even a small quantum of copying may constitute infringement, provided it is the "essential" part of the copyrighted work. Without going into depth about the import of "transformative use", the Court also held that such transformative use might be fair use in some cases. Interestingly, the Court also held that the motive of the alleged infringer is also a material fact in deciding fair dealing.[106]

The Delhi High Court in Chancellor Masters & Scholars of University of Oxford v. Narendera Publishing House,[107] dealt with the nature of "transformative use" in the Indian context. The judgment discusses and heavily draws from the American Four-Factor Test.[108] While expounding the meaning of transformative use, the Court held that a use must be such that it "*serves a substantially different purpose*" from the earlier work, and such use must not be a substitute with minor changes. It must also not affect the market for the earlier work.[109]

TV Today Network Ltd. v. News Laundry Media (P) Ltd.,[110] provides another example of the concept of "transformative use" by holding that comments superimposed on the news excerpts authored by the plaintiffs added value to the said excerpts and since the comments were added with an intent to remove bias from the plaintiff's journalism, it was in public interest and transformative.[111]

---

[102] *Cartoon Network LP, LLLP v CSC Holdings, Inc* 536 F 3d 121, 2008 US App LEXIS 16458, 87 USPQ2D (BNA) 1641, Copy L Rep (CCH) P29,598, 36 Media L Rep 2185, 45 Comm Reg (P & F) 989 (2d Cir NY August 4, 2008).
[103] *Super Cassettes Industries Ltd v Hamar Television Network (P) Ltd* 2010 SCC OnLine Del 2086.
[104] ibid 11.
[105] *Harper & Row Publishers v Nation Enter* 471 US 566 (1985), 548-49, 564-66.
[106] *Super Cassettes Industries Ltd v Hamar Television Network (P) Ltd* 2010 SCC OnLine Del 2086 [8].
[107] *Chancellor Masters & Scholars of University of Oxford v Narendera Publishing House* 2008 SCC OnLine Del 1058.
[108] ibid 1068.
[109] ibid 1086, 1087.
[110] *TV Today Network Ltd v News Laundry Media (P) Ltd* (2022) 5 HCC (Del) 6.
[111] ibid 69.

Given the stringent pre-conditions for invoking the fair dealing defences under Indian law, and the nature of the datasets used and the processing involved, it is highly unlikely that such a defence will be successfully invoked.

# RECOMMENDATIONS FOR BALANCING COPYRIGHT CLAIMS AND PROMOTING TECHNOLOGICAL DEVELOPMENT

## LEGISLATIVE INTERVENTION

It is strongly recommended that swift legislative measures be taken to resolve the current discord in the rights in copyrighted works and the goal of promoting technological development. It is not reasonable to rely on judicial intervention, since such a remedy is dependent on plaintiffs bringing claims in respect of defined disputes, and is unreasonably lengthy in resolving such disputes, owing to technical complexities, pendency in existing litigation, and other such issues with judicial intervention where a quick response is required.[112] Self-regulation is also not a desirable alternative at this stage, since no foundational regulations exist.[113] Lawmakers must focus on the viability of a fair use provision as opposed to a specific licensing regime, the law on hyperlinking, mandating disclosure requirements, and grievance redressal mechanisms while legislating on the issue. The existing law does not adequately address the various challenges that generative models bring forth, and as such, legislative interventions specifically addressing generative models are urgently required. The dispute within the board of OpenAI, is an example of the chaos that self-regulation without effective legislative intervention can bring about, with founder and Chief Executive Officer Sam Altman being fired[114] over alleged differences over whether OpenAI should switch to a for-profit model and on AI safety after stifled rumours suggesting OpenAI had achieved artificial general intelligence surfaced.[115] Creation of a framework within rights management societies to allow quick and easy bulk licensing for training generative models.

---

[112] Communications and Digital Committee (House of Lords), 'Large Language Models and Generative AI' (2024) HL Paper 54, [242]-[247] <www.publications.parliament.uk/pa/ld5804/ldselect/ldcomm/54/54.pdf> accessed 10 July 2025.

[113] Eugenia Lostri, Alan Z Rozenshtein and Chinmayi Sharma, 'The Chaos at OpenAI Is a Death Knell for AI Self-Regulation' (*Lawfare*, 28 November 2023) <www.lawfaremedia.org/article/the-chaos-at-openai-is-a-death-knell-for-ai-self-regulation> accessed 22 February 2024.

[114] Alind Chauhan, 'OpenAI Saga: Sam Altman's Firing, Return, and Future' (*The Indian Express*, 24 November 2023) <www.indianexpress.com/article/explained/explained-global/openai-sam-altman-return-9041014/> accessed 22 February 2024.

[115] Anna Tong, Jeffrey Dastin and Krystal Hu, 'OpenAI Researchers Warned Board of AI Breakthrough Ahead of CEO Ouster, Sources Say' (*Reuters*, 23 November 2023) <www.reuters.com/technology/sam-altmans-ouster-openai-was-precipitated-by-letter-board-about-ai-breakthrough-2023-11-22/> accessed 22 February 2024.

## INCORPORATING A LICENSING REGIME FOR TRAINING OF GENERATIVE MODELS

Copyright societies are recognised under Section 33 of the Act, 1957. These societies function as the nodal point for managing copyrights, granting licenses, managing agreements, and collecting royalties on behalf of authors of works. It is recommended that within the existing framework, copyright management societies be used for the purpose of various copyrighted works, to allow for seamless and quick licensing of works to developers of generative models, such that a wide variety of high-quality data can be provided while still respecting the rights of authors.[116] A fair and accessible licensing framework can contain clauses as to permissible uses, restrictions, payment of consideration, which may be subsidised at a rate which authors and developers agree upon so as to balance interests, and enforcement. The restrictions clause can incorporate governance norms regarding ethical use, can impose duties on developers to adhere to conditions which disallow human impersonation, propaganda, damage to reputation, etc. Thus, a license can also serve as a regulatory and governance instrument.[117] A successful example of using licensing is Reddit, which in February 2024, entered into a licensing agreement with Google to allow training of Google's generative models using user-generated content hosted on Reddit.[118] Reddit, in its Initial Public Offer documents, further claimed that it had made about $203 million out of the licensing proceedings.[119] This goes on to show that licensing can be a viable option for balancing the competing interests of copyright owners and developers of generative models.

The Alternative Compensation System can also act as a balancing act by providing rights holders compensation without bankrupting developers of generative models. A mechanism which estimates a fair value of the training data can be arrived at, and a percentage of the

---

[116] Gregory Smith, 'Licensing Frontier AI Development: Legal Considerations and Best Practices' (*Lawfare*, 3 January 2024) <www.lawfaremedia.org/article/licensing-frontier-ai-development-legal-considerations-and-best-practices> accessed 22 February 2024.

[117] Danish Contractor and others, 'Behavioral Use Licensing for Responsible AI' (2022) ACM FAccT 778 <www.arxiv.org/abs/2011.03116> accessed 22 February 2024.

[118] Anna Tong, Echo Wang and Martin Coulter, 'Exclusive: Reddit in AI Content Licensing Deal with Google' (*Reuters*, 22 February 2024) <www.reuters.com/technology/reddit-ai-content-licensing-deal-with-google-sources-say-2024-02-22/> accessed 24 February 2024.

[119] Kyle Wiggers, 'Reddit Says It's Made $203M so Far Licensing Its Data' (*TechCrunch,* 22 February 2024) <www.techcrunch.com/2024/02/22/reddit-says-its-made-203m-so-far-licensing-its-data/> accessed 24 February 2024).

revenue earned by the outputs of a generative model can be earmarked for rightsholders, who can then share in that revenue pro rata.[120]

## PUBLICATION OF PROCESS OF TRAINING OF MODELS AND DATA PROVENANCE

Developers ought to be mandated to disclose the sources of the training data, and the manner in which it was compiled and used. They must further be obligated to disclose the manner in which the training was done. This would aid in introducing transparency to an otherwise opaque mechanism that are generative models. To balance these disclosure mandates with antitrust concerns and to preserve competition, only that information which would allow courts and regulators to understand the functioning of the model, while allowing trade secrets to be kept confidential, can serve as a viable alternative. Furthermore, developers must make the datasets publicly available to allow the public to search these databases in order to detect their works which are used for training without consent. A grievance redressal body must be set up to ensure the redressal of such claims.[121] Developers can also be nudged, including by legislation, to incorporate privacy by design principles.[122]

However, potential roadblocks to the effective implementation of such publication exist in the level of detail with which developers of such models must adhere to. For example, Recital 68 to the Directive on Copyright in the Digital Single Market is an extant example of the publication requirement, and states that where online-content sharing platforms investigate copyright infringement claims, they are not required to share precise details of all works stored, but merely an approximate summary of the data. This can cause hardship to claimants in proving that their rights were infringed. [123] The EU AI Act, also similarly mandates sharing summaries of training data in Article 56, among others.

---

[120] Amanda Coelho Della Giustina, 'Fair Compensation for Copyrighted Data Used in AI Training' (2024) Masters Thesis <www.arno.uvt.nl/show.cgi?fid=176944> accessed 10 July 2025.

[121] Shlomit Yanisky-Ravid and Sean K. Hallisey, '"Equality and Privacy by Design": A New Model of Artificial Intelligence Data Transparency via Auditing, Certification, and Safe Harbor Regimes' (2019) 46(2) Fordham Urb L J 473–77 <www.ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=2759&context=ulj> accessed 10 July 2025.

[122] Eric Everson, 'Privacy by Design: Taking Ctrl of Big Data' (2017) 65(1) Clev St L Rev 27 <www.engagedscholarship.csuohio.edu/cgi/viewcontent.cgi?article=3933&context=clevstlrev> accessed 10 July 2025.

[123] Adam Buick, 'Copyright and AI Training Data—Transparency to the Rescue?' (2025) 20(3) Journal of Intellectual Property Law & Practice 190 <www.academic.oup.com/jiplp/article/20/3/182/7922541> accessed 10 July 2025.

### CONTENT VETTING BY DEVELOPERS

Developers of generative models and creators of datasets for training such models must be mandated to vet the content that they provide for training generative models. Specific teams must be set up by both dataset creators and developers, with a narrow mandate of scanning the data for illegal, obscene, or otherwise prohibited material. This becomes especially important for dataset creators, since the LAION database was found to contain child sexual abuse images.[124] Legislative intervention can specify the standards of safety for developers of datasets, and for LLM makers, along with procedural safeguards and reporting requirements.[125]

## CONCLUSION

The paper begins by presenting a basic primer on the functionality of generative artificial intelligence models, and while establishing that these models use mathematical probability and statistics, to train the models to recognise characteristics of images, which are then used to create novel material. Part III of the paper then goes on to discuss lawsuits which make claims in copyright infringement, with a view to understand the claims in them, and draw common threads in them. Part IV observes three major perspectives on training data and copyright law, and concludes that it is incorrect to see generative models as collages, that the
data mining exception is a brittle defence to infringement claims, and that no analogue can be drawn between the learning capacity of human beings and generative models. It further states that data mining, which has been permissible under American fair use law, is not applicable to visual generative models. Thereafter, the paper examines the law on copyright infringement claims in India to hold that claims are mainly found in the right to reproduction, communication to the public and violation of digital rights management information. Under Indian law, the claims most likely to succeed are under reproduction, but only in edge cases, and in almost all cases of storage. Communication to the public remains a tacky and uncertain subject, since the law of hyperlinking is not very clear in India and varies widely internationally. It is also seen that violation of rights management information can be a crucial area of litigation. The paper then moves to identify policy prospects and makes recommendations for governing generative models, with a view to balance the competing interests of technological advancement and rights

---

[124] Pranshu Verma and Drew Harwell, 'Exploitative, Illegal Photos of Children Found In The Data That Trains Some AI' (*The Washington Post*, 20 December 2023) <www.washingtonpost.com/technology/2023/12/20/ai-child-pornography-abuse-photos-laion/> accessed 17 February 2024.

[125] Philipp Hacker, 'A Legal Framework for AI Training Data—from First Principles to the Artificial Intelligence Act' (2021) 13(2) Law, Innovation and Technology 257 <www.tandfonline.com/doi/full/10.1080/17579961.2021.1977219> accessed 10 July 2025.

vesting in copyrighted works. Mandating transparency in the use of training data and methods while also balancing competing interests, such as trade secrets, creating an accessible and equitable licensing regime, and content vetting by developers are suggested mechanisms to protect copyright interests. It is hoped that swift regulatory action is undertaken to regulate an otherwise quick developing area of the law.