

Reconceptualising the Concept of Copyright Law with Special Regard to Training of Artificial Intelligence in India

Ms. Simran Chhabra¹, Dr. Tapobrata Pakrashi²

¹Research Scholar, NorthCap University

²Assistant Professor, NorthCap University

Abstract

Artificial Intelligence systems are trained by ingestion of vast datasets of information and computational analysis of copyrighted work without direct involvement of human beings for training of the AI models. These vast pool of information are the very foundation and backbone of functioning of these AI models. However, much of this data is copyrighted and used without consent of the owner. The question as to the legality of such actions gives rise to uncertainty with respect to Copyright Law which was created in the analog era and has no explicit provisions for AI in most of the countries especially India. This technological leap gives rise to the question whether training AI on copyrighted material would lead to infringement of copyright law. This study highlights the ambiguities in the current legal framework and the urgent need for legal reforms. This study would highlight the challenges posed by training of AI models under the Copyright Act 1957, as the conventional copyright law was developed to regulate the expressive exploitation of creative work. Whereas, training of AI models involves computational analysis of copyrighted work rather than expressive exploitation of the work such as adaptation, distribution, communication and reproduction. As AI works on large datasets, including the copyrighted material, it's also imperative to find a striking balance between protecting creators' rights and encouraging innovation and creativity. The study also aims to specifically investigate whether the works fed to the algorithms for statistical learning and training rather than for expressive use falls within the scope of section 14 of copyright act within existing exclusive right or if it can be categorised as a new category of use such as non-expressive computational use. The study aims to explore the legal challenges posed by technological advancements and propose a revised conceptual framework for addressing training of AI within copyright law in India while balancing the interest of the creators.

Keywords: Artificial Intelligence; Copyright Law; Text Data Mining; Infringement; Fair use

Introduction

The emergence of AI has now evolved from performing limited tasks to now performing complex tasks imitating the human brain including answering the queries, formulating answers, creating images, writing music etc. This transformation has not only been achieved by teaching or programming machines to create content but feeding them with vast databases and repositories of human made content.¹ This technological process raises significant question that whether the use of AI training on such material constitute infringement of copyright as traditional copyright law enacted in analog era was developed to protect the works primarily used for human

consumption. Humans saw, read, listened to or experienced the expressive work, whereas the work AI system is fed with for training is processed as data for extracting the statistical patterns, algorithm learning, performing calculations and to extract patterns by converting the data into mathematical reputations, vectors etc. the system only extracts the computational pattern to understand the algorithm and not the expressive meaning of the work. For Eg. A human being would read an article or listen to the song and enjoy it, while AI would read thousands of articles and go through thousands of songs to understand and learn. The language, the rhythm, the pattern, etc. Therefore, it can be gauged as a possibility that AI uses non expressive part of the data for computational analysis than exploit the

¹ Buick, A. Copyright and AI training data—transparency to the rescue? *Journal of Intellectual Property Law & Practice*, 20(3), 182–192. (2025)

traditional expressive copyright law. Another aspect to this could be interpretation of the existing copyright law in traditional sense. Indian Copyright Law has no explicit provisions for AI.² This technological leap leads to the question whether training AI on copyrighted material would amount to infringement of copyright law or can it fall within the scope of section 14 under the existing exclusive rights or if it can be categorised as a new category of right related to non expressive computational right. Section 14 of the Copyright Act grants copyright owner the exclusive right of reproduction, communication, adaptation etc. Feeding AI models with copyrighted content without authorization for training purposes involves creating multiple copies and using content for commercial use without the consent of the author. This action can be considered as transformative use if copies are not stored in databases of the AI companies, otherwise this could amount to infringement of copyright as it would fall under violation of reproduction rights provided under section 14. Section 51 provides for infringement of copyright. The Delhi High Court is currently examining this stance in the case of ANI v. OpenAI (2024). There is also a possibility to cover the act of feeding AI on copyrighted content under the exception of Fair Dealing under Section 52. Many countries like USA have given broader interpretation to this concept to cover this action as fair use. But India's stand on this is still unclear and India usually takes narrower approach regarding the interpretation of the fair use doctrine.

The major conceptual issue is whether computational use of copyrighted work for training AI models be treated in the same way as traditional expressive use of the copyrighted work where the work is actually appreciated and experienced by the humans in contrast to the use to

train AI models where the creative expression is not consumed by the models in a way they are consumed by humans. Rather it is only converted to mathematical units, tokens to unsustainable the statistical patterns and algorithms. This study argues that training AI on copyrighted content should be considered as computational use rather than exploiter in of expressive use and applying traditional laws and doctrines to govern the challenges of new technology would lead to doctrinal and conceptual issues. The study aims to proposes reconceptualisation of the copyright law in the context training artificial intelligence on copyrighted content. It further aims to explore the possibility of categorising computational use working the framework of Section 14 exploring category of uses or added a new category.

AI Training and Copyright Infringement: Understanding the Legal Issues

The growth of generative artificial intelligence (AI) has marked a revolutionary moment in transforming how machines interact with human intelligence and creativity. AI is no longer limited to performing simple tasks, AI now possesses the ability to write poetry, compose music, generate images, mimic human dialogue, and even craft narratives. This has been made possible not just by teaching machines "how to create or answer" in a traditional sense, but by training them on vast databases and repositories of human-made content like articles, poems, books, artwork, films, codebases, and more.³ These massive databases form the backbone of the algorithms on which models like Gemini, ChatGPT, Midjourney, etc. work. In most cases, these databases do not provide knowledge directly to the softwares just out of thin air instead these AI models take data, absorb, process, and emulate patterns from the materials they have been exposed to.⁴

² Seshadri N. The Dilemma of Artificial Intelligence Generated Works and Indian Copyright Law. Indian JL & Legal Research. (2021) Available at: <https://www.ijlrr.com/post/the-dilemma-of-artificial-intelligence-generated-works-and-indian-copyright-law>

³ Pescapè, A., 'Exploring the Current State and Future Potential of Generative Artificial Intelligence Using a Generative Artificial Intelligence.' In:

Santoian, F., Giannini, G., Ciasullo, A. (eds) Mind, Body, and Digital Brains. Integrated Science, vol 20. Springer, Cham. (2024)

Available at: https://doi.org/10.1007/978-3-031-58363-6_4

⁴ Paquette, L. (n.d.). 'Artificial life imitating art imitating life: Copyright ownership in AI-generated

In this context, data becomes the fuel to run these AI models and much of this fuel is copyrighted. Generative AI models rely on ingesting, scraping, and analyzing large volumes of data often sourced from the open web without any permissions or licenses. This includes copyrighted books, songs, novels, journals, research papers, paintings, news articles, etc. While these materials give the AI the data and context it needs to produce "human-like" results, their constant use has triggered a backlash from creators, artists, writers, and publishers across the world. The present day AI models are majorly dependent on already existing intellectual works, which raises questions about legality, ethics, and compensation in the training and developing stage itself. Adding to this technological complexity is the fact that the results of generative AI are presented as original works, while their foundation lies buried within the existing data and preexisting creations. The AI models can said to be inherently derivative, and their value directly depends on the quality and diversity of the data they are trained on. This raises a the issue whether AI models owe their competence to the available data and copyrighted content, and should creators have a say or a share in that process.

The dilemma is not whether AI can or should learn from human-created material, it surely can and already does. But the real concern is whether such learning is ethical, lawful, and sustainable. For India, this presents a new legal dilemma where the law must develop with the rapid growth in innovations that has already crossed many boundaries. The current dilemma provides risks harming both the tech industry and the creative ecosystem. The major concern regarding this issue is whether using this data to train AI models should be considered using of copyright in traditional expressive sense or could it be considered computational use. And in case we consider it as computational use in modern sense as a necessity for development of technology, would it amount to

infringement of copyright or can be covered under the ambit of Section 14.

TRADITIONAL FRAMEWORK OF COPYRIGHT ACT 1957

The evolution of copyright law in India dates back to 1914. The Copyright Act of 1914 governed copyright protection before the Act of 1957 which was an expansion of the British Copyright Act of 1911. The 1957 Act was amendmended various times specifically in 1983, 1984, 1992, 1994, 1999 and 2012. The Copyright Amendment 2012 brought Indian copyright law into accordance with the WCT and the WPPT, two treaties of the World Intellectual Property Organization.⁵ This Amendment Act of 2012 modified the rights of sound recordings, motion pictures, and artistic works by declaring that "reproducing" an artistic work, making a copy of a cinematograph film, or embodying a sound recording now includes "storing" it in any media by electronic or other means. This contributes to addressing some of the challenges that develop in the contemporary digital age.

It also provided for "commercial rental" rights for computer programs and cinematograph films as

Mentioned in Article 11 of the TRIPS Agreement, Article 7 of WCT, and Article 9 of WPPT. The act outlines the rights granted to the copyright owner of literary, dramatic, musical, or artistic works, cinematograph films, sound recordings, and computer programs, as described in section 14. Additionally, this provision provides many rights to safeguard the interests of broadcasters, entertainers, and others. To safeguard the interests of actors, dancers, musicians, and others, the Copyright Amendment Act of 1994 added clauses about "Performers' rights" in Section 38. Copyright focuses on the exclusive right provided by law to the creator of the work. Authorship refers to the person who creates the book, who is also the owner of that particular book. Authors, or those who create

works.⁷ Available
at:<https://www.proquest.com/openview/5f506b6738a40a3833621f532ccf6310/1?pq-origsite=gscholar&cbl=46743>

⁵ Sharma, Ananya, Artificial Intelligence And Copyright Laws: A Critical Analysis Of

Challenges And Legal Adaptations (April 28, 2025). Available at SSRN: <https://ssrn.com/abstract=6159606> or <http://dx.doi.org/10.2139/ssrn.6159606>

intellectual property in the form of sound recordings, cinematograph films, and literary, musical, dramatic, and creative works, are protected by copyright. The first owner of a work's copyright is often the author. This rule got international recognition through the Berne Convention. It is also the international norm that there is no need to satisfy any formalities, such as registration, for the author to enjoy the rights. However, the Berne Convention does not define the term "author". Instead, it provides that in the absence of proof to the contrary, the person under whose name the work is disclosed is to be regarded as the author.⁶

Technological revolution has completely evolved the way creative works are shared and consumed in the digital age and this has given rise to many legal challenges. In today's world, it has become much easier to copy and distribute content, leading to copyright infringement and piracy. To tackle this difficulty an attempt was made by bringing certain amendments to copyright law by WIPO in the form of WCT AND WPPT. The WIPO Copyright Treaty (WCT), which was established in 1996, expands on the fundamental ideas of the Berne Convention and attempts to ease the difficulties of the digital world. The WCT recognizes the need to modify copyright laws to reflect contemporary digital realities by extending copyright protection to digital works disseminated online. DRM systems, regulate the access and use of digital content, are protected by clauses in this treaty. The WCT also has safeguards against technological protection systems being circumvented, strengthening copyright protections in a time of swift technological advancement. In the digital age, copyright refers to the protection of

digital works, which is necessary to ensure that authors receive fair compensation for their labor and to promote creativity and innovation in the digital sphere. Currently, there are several ethical and legal issues surrounding the protection of digital works created by artificial intelligence (AI). Regarding copyright, for instance, there are issues with identifying and safeguarding digital works produced solely by artificial intelligence. Since AI works are linked to a human creator, they are not protected by copyright under present laws in the majority of countries. This brings up questions about how human creators who work to create and train AI algorithms that produce such things should be acknowledged and compensated.⁷

AI TRAINING AND USE OF COPYRIGHTED CONTENT

AI is concerned with machines demonstrating the intelligence of a human mind and makes them capable to imitate the learning and decision making abilities of a human mind.⁸ Artificial Intelligence is trained on vast amount of data repositories. Machine learning involves learning from the data fed, improving the performance, giving predictions and making decisions based on input. AI model are trained to understand the patterns in data, and give decisions accordingly. AI is trained on vast amount of data. In layman terms, it is done with the help of machine learning which involves learning through the data fed for training the models.⁹ AI models are programmed to understand the patterns in data, make predictions and respond accordingly. This data comprising of images, text, audio's, videos etc. can be used for supervised, unsupervised and reinforced learning, programming algorithms, evaluation,

⁶ N.S. Gopalakrishnan and T.G. Agitha, Principles of Intellectual Property 305 (EBC Publishing (P) Ltd.,34-A, Lalbagh, Lucknow, 2014).

⁷ Ananya Sharma, 'Artificial Intelligence And Copyright Laws: A Critical Analysis Of Challenges And Legal Adaptations' (2026) Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6159606

⁸ Andy Peart, John McCarthy, 'The Father of AI'

Available

at:<https://www.artificialsolutions.com/blog/homage-to-john-mccarthy-the-father-of-artificial-intelligence>(Visited on 7th April,2026)

⁹ Dr. Sonali S. Gholve, Adv. Sagar K. Gholve, Study of whether using Copyrighted Content for Training AI Models a Copyright Infringement?,"International Journal of Humanities, Commerce and Education,ISSN:3108-0456 (Online), Volume 1, Issue 3, pp. 25-33, November 2025.

Available

at: <https://ijhce.com/index.php/ijhce/article/view/14/14>

continuous learning etc.¹⁰ However, while training AI models certain challenges are encountered, i.e. incomplete and distorted data produces inaccurate results, interpretation of the result can vary as rationale or logic Behind the decision given by AI is unknown, training of AI models also involves high cost and sometimes it is only able to respond to the data that has already been fed to it and unable to cater to other queries.

Copyright protection means legal rights given to creators of work like music, book, films, software etc. The data which gathered in a unique and original way such as databases can be considered as copyrighted data. But raw data or raw facts are not considered copyrighted. Main legal concerns of using AI and copyright are to use copyrighted material to train AI models and second is work generated through this process. Lack of clarity regarding the rights of content creation and need to redefine the scope. Use of copyrighted material to train AI is a major problem. According to AI company its fair use is like human minds reading the data while owners of that data argue that its infringement. But balancing between user creators is still unsettled.¹¹ Training of AI models on copyrighted vast amount of data available online including the copyrighted content raises several issues pertaining to copyright law.¹² Some scholars argue that this usage of copyrighted content for training purposes without authorisation by the owner leads to copyright infringement as machine learning uses unauthorised input datasets for training purposes and in certain cases also makes digital copies of the work fed to them for future assistance and non literal reproduction in models during the input stage.¹³ Once such data is fed to the

AI models, there are high chances that the output would also be similar to the input fed. So, at this stage also there is a possibility of copyright infringement. Therefore, there is a possibility of copyright infringement at all 3 stages, input stage, during the training process and output stage.¹⁴ However, this action of training AI could also be understood as a necessity for technological development where the intention of the developer is not to infringe or hamper the economic rights of the copyright owner but only to contribute towards the technological development. And to further this view, it could be considered that the copyrighted material was only used for non computational or transactional use where it was not stored or consumed for entertainment but only for understanding patterns, extracting algorithms, and learning the statistical models for learning purposes. So, this could fall under the exception of Fair Use as interpreted by courts of some countries by taking liberal interpretation of the doctrine of Fair Use. Another possibility could be providing this as an exclusive right to be granted under Section 14 for the use of copyrighted content for computational use by the copyright owner where the copyright owner can be sufficiently rewarded for his creativity and the AI developers could get license or access to the data. This necessitates the need for amendment to the present copyright law.

Challenges under Copyright Act

AI models are trained using vast amounts of data including books, music, articles, images, artwork, etc that is most certainly protected by copyright. To programme an AI model, developing companies usually acquire, copy and store this data in different formats. Even if the copied data isn't shared on

¹⁰ *ibid.*

¹¹ Dr. Sonali S. Gholve, Adv. Sagar K. Gholve, Study of whether using Copyrighted Content for Training AI Models a Copyright Infringement?", International Journal of Humanities, Commerce and Education, ISSN:3108-0456 (Online), Volume 1, Issue 3, pp. 25-33, November 2025.

Available at:
<https://ijhce.com/index.php/ijhce/article/view/14/14>

¹² Sujith Ravi, "On-Device Machine Intelligence", Google Research Blog (Feb. 9, 2017) Available at: <https://perma.cc/WQ8L-WS5D> (visited on April. 23, 2026)

¹³ Brendan McMahan & Daniel Ramage, "Federated Learning: Collaborative Machine Learning without Centralized Training Data", Available at <https://perma.cc/XVA2-J96J>

¹⁴ Andres Guadamuz, "Artificial Intelligence and Copyright", 23 WIPO MAG (2017) Available at: <http://www.wipo.int/tipso> (Last visited on 23rd April, 2026)

public platforms, acquiring, storing and using it to train AI without permission might amount to infringement of copyright laws. Courts in many countries have established that even keeping copyrighted data on a company's private server, without sharing in any public domain can still amount to copyright infringement. Furthermore, most countries still don't have any specific regulation about whether using of copyrighted material to train AI is legally permissible or not, it's still falls under a legal grey area. Some scholars argue that using copyrighted data in this without permission would violate copyright laws unless there's a specific valid exception like "fair use", which has been established in few countries for certain type of copyrighted material like USA and Japan.

It is argued by the tech companies that training AI on copyrighted data would count as 'transformative' that is changing the material enough to count it as new data. It is argued that AI doesn't store data in exact format but instead learns algorithms based on patterns, which helps it later to provide information and create new things as per user requirement. For example, AI may learn the usual way or format a novel is written or how a certain type of artwork looks and then create or provide its own versions. This is argued to be different from copying and reproducing the original content or work. But on the other hand others argue that the AI is trained to produce and create the content on the basis of the work that is copyrighted and used without permission to train AI which is violative of the legal norms and leads to serious copyright issues.

Adding to this fact is the concern that AI Tools sometimes memorises some part of the data they are trained with, which increases the possibility of similarity in output. This means that AI has the tendency to produce exact images or repeated sentences from the training data, which again amounts to copyright infringement despite the fact that training of AI was done within or as per the permissible legal norms. Many large AI companies are working on this issue to make sure that the output generated by AI doesn't amount to copyright infringement by scanning and filtering the data before responding.

This is where legal exceptions like "fair use" (in the U.S.) or "fair dealing" (in countries like India and the UK) come into play. In the U.S., the law looks at four main factors to decide if something counts as fair use: (1) why the work is being used, (2) what kind of work it is, (3) how much of it is being used, and (4) whether the original work's market is harmed. Courts may allow the data to be used for training if the data that is factual or publicly available is used for research, and this new use doesn't affect the market for the original work. But the data used for training AI includes as it is copyrighted books or movies to create tools for profit of the tech companies, then it could be considered copyright infringement. So there isn't one clear straightjacket solution, rather it depends on how and why the AI was trained or how the data was used.

International Approach

The question as to whether AI developers can legally use copyrighted content for training their models is addressed differently in all countries. Under India's current copyright law this question is both complicated and largely unsettled. AI systems like Gemini, ChatGPT or image generators work by processing and training from vast amounts of content including text, images, and other expressive material, much of which is protected by copyright. The procedure of analysing, taking, storing, and using this material during AI training may, *prima facie*, amount to reproduction and therefore, copyright infringement under Section 51 of the Indian Copyright Act, 1957. Indian copyright law allows for the exception of "fair dealing" (Section 52) under copyright law, but this concept is very narrowly defined and interpreted by the courts. The exception of fair use can include data used for personal use, research, educational purpose, reporting and criticism. It does not extend to broad or commercial AI applications. Fair dealing in India was never designed to cover data mining or algorithmic scraping at a scale it used in AI training.

Some researchers have argued that if the content generated by AI is significantly different from the original work, it might qualify as transformative. As 'transformative' refers to the creation of a new content, expression, message, or meaning that

significantly alters/changes the original work's character or purpose. This concept is not expressly mentioned in Indian copyright law. While Indian courts have mentioned this concept in the past cases, such as *Civic Chandran v. Ammini Amma*. However, there is no concrete precedent applying this concept to AI models or datasets. This argument holds very limited weight because potential copyright infringement arises primarily at the input stage, where copyrighted content is directly fed to AI and reproduced. Even if the final AI generated result doesn't resemble any original work, the training and programming phase itself can still be violative of the copyright law. Especially in the case where training material was taken without permission or by commercial setting. Courts may also take into consideration if such use harms the economic interest or market value of the original output something that still falls in a grey area.

Furthermore, India lacks specific legal provisions for text and data mining (TDM), that has been introduced by few jurisdictions such as the Japan and European Union, where such cases are addressed through explicit exceptions. This ambiguity in India's legal framework creates gives rise to uncertainty for creators and technology companies, working towards development and innovation, also hindering international collaborations at the same time. The absence of explicit TDM exemptions under the Indian Copyright Act, 1957, as provided under other jurisdictions such as EU, forces AI developers to negotiate and enter into complex licensing arrangements, which acts as a barrier for small enterprises and researchers who lack the resources to secure such permissions as it required a lot of funds and R&D. Moreover, AI developers in India working without authorisation and licenses may have to suffer legal implications and sanctions, especially when their models are commercialised for using copyrighted content without authorisation. European Union on the other hand, provides for the

European Union's Directives on Copyright i.e. Text Data Mining Exception (TDM), fostering a more flexible system for facilitating AI advancement.¹⁵ Similarly, Japan's 2018 Amendment under Copyright Act 2018 explicitly provides for the regulation of AI development and data analysis through relevant exceptions. Without adapting to the technological advancement, Indian developers are bound to face a risk of infringing copyright laws while training AI models which could lead to litigation and financial liabilities. This legal uncertainty not only restricts growth, but also affects India's competitiveness regarding the technological development in the global AI landscape.

Development of technological models such as AI is essential for technological progress, but the system of using copyrighted content and datasets without consent of the owner is not protected under the current Indian law. There is a need for the legislature to explicitly update the Copyright Act 1957 or courts should set clear precedent with respect to the use of copyrighted material in training or programming the AI models particularly for commercial purposes that would lead toward infringement and not fair use.

Different countries handle the issue of training of AI models and regard to copyright law in different ways. In the United States, the Copyright Act (Section 107) allows "fair use" exception. U.S. courts have expanded their horizon and started applying this idea to AI training, especially when the training changes the material in a meaningful way with different output altogether from the input fed and serves a public or research purpose. Ongoing cases like *Andersen v. Stability AI*¹⁶ and *Authors Guild v. OpenAI*¹⁷ are shaping how this plays out in practice.

In India, the Copyright Act of 1957 does not clearly talk about how AI can use data for training. There are some rules under "fair dealing" that allow limited use of content, but it's not clear if these rules apply to the large amounts of data AI needs. This

¹⁵ Lucchi, N. (2025). Generative AI and copyright: Training, creation, regulation. European Parliament. Available at: <https://repositori.upf.edu/server/api/core/bitstreams/9a3551f8-0d83-4a02-a4b4-6bea3b121a02/content>

¹⁶ *Anderson v. Stability AI Ltd.* (2024), 2024 WL 3823234

¹⁷ *Authors Guild v. OpenAI Inc.* (2023)

confusion shows that India needs new laws to manage how AI uses content, making sure both creativity and copyright are protected.

But in the United States, courts have already seen lawsuits against companies like Anthropic and Meta, which used copyrighted books to train their AI tools. In the case of *Authors Guild v. Anthropic*, the court said that if the books were legally bought or scanned and the AI was trained for non-commercial reasons, then it could fall under “fair use”. But using pirated or scraped data was seen differently courts are stricter when content is taken from illegal sources. Judges also made it clear that people who claim their work was copied need to show that the AI model produced outputs very similar to their original work. So even if training itself might be legal, the AI’s output could still break copyright rules if it closely matches the original.

The Copyright Law in UK provides for exception in the case of non-commercial research by providing for “text and data mining” for exception. This means that if AI is being trained for commercial or business purposes, companies have to get a authorisation in the form of license or permission from the copyright owner. To reward the creator sufficiently and work towards betterment of the society.

The European Union has issued EU Directive of 2019 which provides for allowing research groups to do text and data mining under Article 3. Article 4 allows anybody to use data in a way that is only for research purpose and not for final consumption unless the copyright owner has clearly said they don’t allow it. However, many rights holders now also provide for opt-out notices to block their works from being used in AI training in which the developers have to respect their choice and not use their data in any case and explicitly refrained through opt out option.

Japan’s copyright law is one of the most AI-friendly copyright laws. Article 30(4) of Japan’s Copyright Act allows people use material for machine learning that is copyrighted, as long as it’s not being used for

people to read, watch, or listen to directly. This gives developers more freedom, even for commercial AI projects.

India, doesn’t have any specific law for AI training yet. Section 52 of the Indian Copyright Act lists some “fair dealing” exceptions such as for personal use, private research, or education but these are limited and unclear. It’s not certain whether AI training counts as “research” under this law, especially for companies. Experts say India needs new rules to deal with these modern challenges.¹⁸

Reconceptualising of Copyright Law for AI Training

The ever evolving technology of artificial intelligence necessitates the reconceptualising of copyright law, specifically in the area of training of AI models on copyrighted content. Traditional copyright laws were designed in the analog era to regulate the expressive use of the works such as broadcasting, reproducing, distributing, communicating etc., where the work was ultimately consumed by the human beings. Whereas AI training represents a different paradigm which involves non computational use of the work for training purposes including ingestion of large databases, conversion of data into statistical patterns for learning. So, in the latter, the data is not used for ultimate enjoyment by the humans but only for extraction of patterns, statistics, correlations etc. So, generalising all such works as copyright infringement straightaway would cause hinderance the technological development. So, a balanced approach should be adopted balancing both innovation and development in technology and safeguarding of creator’s interest. Such use can be considered transformative use or can be allowed as a separate category under Section 14, as exclusive right granted by the authors, for which authors should be rewarded as well for their work.

Conclusion

The rapid growth of Artificial Intelligence has given rise to many questions regarding the basic structure

¹⁸ Leibler, Y. ‘From infringement to innovation: Reimagining copyright for AI training datasets.’(2024)

of the copyright law with regard to regulation of Artificial Intelligence and its training, particularly under the Copyright Act, 1957. The traditional law, was designed to regulate the the aspect of expressive use of creative works through the rights granted for reproduction, adaptation, distribution, and communication to the public as per Section 14 of the Act. However, the process of training AI systems poses different aspect of using copyrighted content.by introducing different kind of interaction with copyrighted material, such as using the material for non-expressive, statistical, and computational purpose. This paper explored that the reading, ingesting and analysing of copyrighted works for the purpose of training AI models cannot be regulated properly within the traditional law under Section 14. The absence of direct expressive output, along with the transformative, and non-consumptive aspect of such use, creates ambiguity for the basic doctrines of copyright laws . Also, the unregulated use of copyrighted material without authorisation from the authors raises legitimate doctrinal concerns regarding the infringement of authors 'rights and economic interests.

The current legal framework in India neither explicitly permits nor clearly prohibits such uses, thereby creating uncertainty for both rights holders and AI developers. This ambiguity risks stifling innovation on one hand and undermining the incentive structure of copyright law on the other. The study therefore concludes that a strict application of traditional copyright principles to AI training is inadequate and conceptually strained. Accordingly, there is a compelling need to reconceptualise copyright law to accommodate technological realities. The recognition of AI training as a distinct category of use separate from expressive exploitation emerges as a necessary doctrinal shift. Such a reconceptualisation must be guided by a balanced approach that preserves the core objectives of copyright protection while enabling innovation in the field of artificial intelligence.

Recommendations

1. Recognition of Non-Expressive Use under Indian Copyright Law

The legislature should explicitly recognise “non-expressive computational use” within the Copyright Act, 1957. This can be incorporated under Section 14 to be granted as an exclusive right by the creators. This would help regulate the use of copyrighted works for AI training, where the final output is not same or where there is no direct expressive reproduction or substitution of the original work.

2. Introducing Text Data Mining Exception for AI Training

Text Data Mining exception should be introduced In Indian Copyright law, permitting the use of copyrighted works for AI training under defined conditions. This exception may draw inspiration from text and data mining (TDM) exceptions in comparative jurisdictions Like European Union, while being adapted to the Indian context while taking into consideration factors such as lawful access, purpose limitation, and non-substitutability.

3. Providing Licensing Framework for Training Purposes

For the purpose of training AI Data, a compulsory or collective licensing mechanism may be introduced to ensure the balance of technological innovation and rewarding the creators for their innovations.. This would ensure that authors are compensated for the use of their works while avoiding transaction costs and practical barriers associated with obtaining individual permissions for large datasets.

4. Expanding the Scope of Section 14 -Exclusive Rights

Clarification is required to determine whether acts involved in AI training such as temporary copying, data extraction, and pattern recognition fall within the scope of “reproduction” under Section 14 or a separate category of non transitional use can be incorporated under the scope of Section 14.

5. Incorporation of Fair Use Doctrine

The fair use doctrine under Indian law in very narrowly interpreted unlike other nations like US and Japan and could be expanded to incorporate a more flexible approach similar to fair use. This would allow courts to govern AI training practices

on a case-by-case basis, based on use, purpose, transformation, market impact and final result.

6. Transparency Requirement and Data Management

AI developers should be subject to obligation for transparency disclosure regarding the datasets used for training. Disclosing the material used for training would promote accountability and credibility and enable creators to enforce their claims.

7. Balancing Innovation and Creator's Rights

A structured balancing be created, to weigh the interests of copyright holders and technological innovation at the same time. It is necessary to

balance both for development of the society and protecting the creative interest of the humanity. Factors such as public interest, transformative value, and economic harm should guide this assessment.

8. Promoting of Open Access and Public Datasets for Published Material

Policy enacted should promote the development and Use of data available on open-access databases and publicly licensed content for AI training. This would reduce dependency on copyrighted material, give due credit to the creators with sufficient reward through licensing and foster innovation within a legally compliant framework.