RESEARCH PAPER

AI Training and Copyright Infringement in India: Issues of Ownership, Consent, and Accountability

[1]Rupam Saikia, [2]Zeeshan Rahman
[1]Chairperson, Centre for Juridical Studies,
Dibrugarh University. Email: rupamsaikia@dibru.ac.in.
[2]Law Graduate, Centre for Juridical Studies,
Dibrugarh University. Email: zeeshanrahman306@gmail.com

ABSTRACT

This research paper examines whether the use of copyrighted works in training generative artificial intelligence systems constitute infringement under the Copyright Act, 1957. We have advanced arguments signifying that AI training necessarily facilitates reproduction and electronic storage of copy righted contents without consent with reference to Sections 13, 14 and 51 of the Act and legally crafted fair dealing exceptions under Section 52. We have critically reveled that such exceptions are not extended to large-scale commercial training practices . The article has further analyzed that the Digital Personal Data Protection Act, 2023, excludes publicly available personal data from the protective shield and has identified the regulatory gap in the governance of AI training to the said effect. Through doctrinal approach and comparative study with United States' Fair Use Doctrine and European Union's Text and Data Mining Regime, the research paper has established a structural mismatch between traditional remedies for infringement of Copy Right in India and for nonconformity to some critical aspects relating to enforcement of Copy Right Act under AI Architecture as evident in emerging rapid developments with the AI pattern . It concludes by proposing consent-based and responsive institutional reforms to reconcile technological development with an ineluctable protective shield for human creativity to facilitate harmonious co- existence of two critical components .

Keywords: Artificial Intelligence, AI Training Data, Copyright Infringement, Fair Dealing & Data Mining.
How to cite this article: Saikia R, Rahman Z. AI Training and Copyright Infringement in India: Issues of Ownership, Consent, and Accountability. Int J Drug Deliv Technol. 2026;16(33s):552-565. DOI: 10.25258/ijddt.16.33s.67

INTRODUCTION

Artificial intelligence (AI) has emerged as the leading technological innovative leap of the twenty first century triggering conspicuous changes in our life . What we sensed as a dream or science fiction at one point of time, has been unfolded in to reality, now. The AI large scale models are trained by the developers on pre-existing data available in public , such as audio recordings, visuals, images, drawings etc. and are collected by the developers. Such collections are "fed" into AI systems which are used by AI regime to identify patterns and generate own responses. This approach of using online data has led to the accelerated growth of AI, but one simple question remains to be answered and that is , who owns the data being used to train AI? It is well established legal principle that, an intellectual creation is owned by the author unless he transfers the right to another **party[1].** It has raised serious legal concerns regarding requirement for consent of authors for using such works as data as well as the liability of companies for using the same. The controversy regarding the use of such data without consent of authors has exacerbated strains in between technological innovations and misappropriation of the creative property of copy right holders for propelling the trajectory of such innovations. For creators, the unauthorized use of their works to train AI is an indication of commercial exploitation as well as assailment of creative autonomy. But for AI developers, such data is of considerable importance for the development of accurate AI systems. This conflict of interests remains to be legally addressed for finding out an

efficacious remedy. It is reasonable to presume that on one side we have to facilitate technological innovations and on the other hand we have also to follow the fundamental principles of Intellectual Property Law for shielding substantive copy rights of the authors coupled with neighbouring rights.

The core legal difficulty lies in determining whether AI training merely learns from copy righted works or whether it reproduces them conforming to mandates of the Copyright Act, 1957.

In this debate, India is saddled in a complex position in contrast to the United States of America(USA or US) which has an open-ended doctrine of fair use or the European Union(EU), which has introduced specific restrictions and exceptions for AI Data Training and mining, India has introduced the Digital Personal Data Protection Act of 2023, which has introduced a consent-based approach for using personal data while excluding data in public domain from the said stipulation . This creates a situation where AI training data may get a protective shield under the Data Protection Law to some extent. However, the question of copyright infringement remains reasonably distinct from the data protection compliance, but at the same time, legal and ethical concerns in regard to protection of personal data and copyrighted creations in AI Regime require to be fairly addressed for harmonious existence of human creativity and technological progress.

This paper aims to focus on the legality of AI training data under copyright law. It proceeds on the premise that at the reproduction stage of AI training, rather than similarity of generated outputs, primary aspect of legal scrutiny should include evident intersections between two inescapable dimensions of Laws. **It proceeds on three key aspects: ownership of data, consent of copy right holders and the liability of the AI developers**. It also attempts to show that Indian copyright law along with data protection law have already laid a foundation of a legal principle capable of regulating AI training Data and insightful analysis would heighten its **importance**[2]. The difficulty, therefore, not stems from doctrinal quiescence, but for lack of responsive structural adaptation. The need of an efficacious legal instrument and mechanism addressing this issue substantially draws attention of the academic world for carving out a harmonious solution to the issue *in-situ* . While we may rely on foreign legal doctrines after critical scrutiny of contexts for addressing such situations, we should also be concerned about their operational limitations in our system. As such, foreign laws are not always easily applicable in India. The analysis that follows situates the AI training within existing legal principles while identifying the limits of its present enforcement architecture. The research questions may be discussed as below:

Research Questions

This research paper focuses on three main questions:

1. Under the Indian Copyright Act, does the use of copyrighted data in AI training amount to infringement and whether exception, such as fair dealing is applicable to the same or not?
2. How does India's Digital Personal Data Protection Act,2023 handle personal Data used in AI training?
3. How does Indian Legal Regime would accept the legal approaches opted by USA and EU in AI training Data?

AI Training and Legal implications of Data Use

Training of AI systems involves the meticulous collection of data and the processing of said data to empower the system to understand the correlation between data points. In other words, this method converts the human created data such as images, visuals, audio, etc, into mathematical representations. But legally speaking, this process of training is highly dependent on the reproduction and storage of protected works. Under the Indian Copyright Act 1957, the concept of reproduction is defined broadly. The Section 14 of the Act protects the exclusive right of copyright holder to reproduce or store any content that he owns. This framework was kept in place keeping in mind the digital copying and storage of **works**[3]. This means that the use of such works being used as datasets in training of AI, even if copied on a temporary basis or used by the developers for analytical use, it violates storage and reproduction **rights.**[4]

AI developers in defense always argue one thing, that the AI models, even if trained on such data do not produce copies of the original works and therefore they do not infringe upon copyrighted contents . However, the Indian copyright law stipulates that , for infringement of a copyrighted work , there is no need for the work to be identical to the original work nor does it require public circulation. The unauthorized act of reproduction in itself is sufficient to establish copyright infringement. Therefore, the legal focus should be on the process of training such AI models, rather than originality or market impact of generated output. Therefore some scholars have argued that, in AI system copyright protection would matter in terms of use and storage

of **copyrighted data.**[5]

An additional defense advanced by the AI developers is that even human beings learn from data. A person learns to draw by watching someone else drawing and he or she is inspired by it and attempts to draw similar things in the beginning. But learning by an AI system is different from that of learning by a human being in the natural process. The AI uses storage systems to store data and then only it is processed. Human beings on the other hand, use their cognitive ability to understand data, it does not involve in storing or reproduction of copyrighted data. Indian courts have stated that law regulates act not metaphors and does not recognize any general right to reproduce works merely for the purpose o**f learning**.[6]

Copyright Protection and Centrality of Human Creativity

Copyright law in India is based on the principle which it completely human centric. It is founded on the idea that protection should be given to intellectual labour and creative expression of creators as laid down in the Copyright Act, which limits protection to "original" literary, dramatic, musical and artistic **work**[7]. Indian courts have many a times stated that "originality" requires application of human skill, judgement, rather than novelty in the mode of expression. The Hon'ble Supreme Court of India clarifies that copyright is practically non-existent in pure mechanical **reproductions**[8]. By rejecting the doctrine of "sweat of the brow" and accepting a new standard that is based on skill and judgment of a human being for creation of a copy righted work , the court in turn reaffirmed that intellectual contributions of a human being is inseparable from copyright protection. The law does not protect the data, but the human contributions within its forms of expression. Human creativity as the integral component for copyright protection is also argued by various **scholars.**[9]

The copyright law further emphasizes on maintaining integrity of authorship of intellectual creations and their protections. It states that the power of authorship shall only reside in natural persons with creative choice & judgment and therefore the Law does not treat computer generated work as work of an independent **author**[10]. So, using copyrighted works to train AI is not just a technical step. It directly affects the rights of the people who created those works. Furthermore, emphasis on retention of integrity of human creativity can be understood from the Section 57 of the Act which recognizes moral rights of authors of intellectual **works**[11]**.** The Delhi **High Court**[12] stated that the harm caused to the work of an author can be interpreted as assailing of honour and reputation of the author. This is done by the AI systems that fragment the original creation of an author and do not even give him or her due acknowledgement, It transforms the authors' works into data points and distorts creative forms & expressions as well as originality of copy righted works. The Indian Copyright law is not framed only for protection of commercial interests of authors, but also it substantively protects creativity and intellectual labour of the author of a work coupled with mandates for maintaining integrity of form and value .

It is concerned with protecting the legal and moral bond between creators and their works. So, when we examine AI training under Law, we should start with a basic idea: copyright is meant to protect human creativity and new technology cannot ignore such protection unless the Law clearly carves out exceptions for it.

Scope of Copyright Protection under Indian Law

Sections 13 and 14 of the Copyright Act, 1957, primarily determine the scope of copyright in India. The Section 13 of the Copy Right Act 1957 mentions about the categories of works in which copyright exists **or is applicable**[13], while Section 14 gives exclusive rights to copyright holder including **neighbouring rights.**[14]These two sections jointly form the legal boundaries within which acts such as copying of works for AI training should be evaluated. The Section 13 limits copyright protection to original literary, dramatic, musical and artistic works as well as cinematography and sound recordings. Once a work crosses or meets the originality threshold set by the law, it is protected against unauthorized use. The data used to train AI are usually literary texts, images, audio recordings, and visual contents that fall under these categories or classes of works being protected under such stipulations. This leads to a reasonable presumption that inclusion of such works in AI training datasets is a clear violation of an author's Intellectual Property Rights against unauthorized use of his creative works as data . The section 14 is central to the question of training AI. It grants the owner exclusive right to reproduce or store the work in any form as he pleases, including electronic form. By clearly including electronic storage of copy righted works by an author , the law has straightened the legal argument that the legislature also intended to cover digital copying as infringement of such a right and therefore there is no ambiguity to opine that

right to digitalization of a work is only delegated to the author or to an entity with his consent or authorization of him. When such works are copied and stored for AI training, even if for analytical purposes, it clearly violates the right of the creator/ author of the work under Section 14 of the Copy Right Act.

Under Indian Copyright Law it is not required that reproduction of a work should be permanent in nature, publicly distributed or commercially exploited for terming it as **infringement of copy right .**[15] The Copyright is infringed as stated in Section 51, even when any act that is exclusively reserved for a copyright owner is done without the owner's authorization. Therefore, this statute focuses on the unauthorized use of such rights rather than on examining similarity of the original onewith the copied work for the purpose of Copy Right infringement. This demonstrates that even if AI models may not fully replicate data, they still infringe on authors' copy rights during the training stage. Moreover sections 13 and 14 do not lay down the difference between expressive and non- expressive use of copyrighted works. The main focus of the statute is to regulate the reproduction and storage of copy righted works . Unless this use falls within an exception as stated in the **Section 52**[16] of the Copy Right Act , unauthorized copying is not permissible . Thus, use of such copyrighted data in AI training can only be justified if permissible under the exceptions of the Section 52 of the Act.

It can be safely presumed that the section 13 and 14 combined together give copyright owners authoritative control over copying and storage, even under the AI regime. We have to decide whether AI training can be placed under legal exceptions in India before concluding that AI Training process can be an exception under Section 52 of the Copy Right Act 1957. The section 52 deals with the fair use doctrine and the Indian legal framework has not endorsed a general fair use rule.

Limitations and Exceptions under Indian Law

Even though Section 13 and 14 can grant broad exclusive rights to copyright holders, under Section 52 of the Copy Right Act, the relevant law has recognized limited exceptions which are referred to as "Fair Dealing". These exceptions lay the foundation for allowing the copyrighted works to be used in certain cases that are listed in the statute and upheld by the courts of law. Unlike the laws of USA that follow an open ended fair use doctrine, the Indian approach to this is specific and delineated. It means that if the use of such copyrighted data does not fall in any of the categories mentioned in the statute, same cannot be justified merely on grounds of fairness or public **interest.**[17] Section 52 of the Act permits the use of copyrighted data or fair dealing for the purpose of research, criticism, reporting of events etc. AI developers often attempt to argue that use of copyrighted data falls under fair dealing as they use such data for "research", taking shelter under Section 52(1)(a) of the Act. However, Indian courts have interpreted 'research' in this context as to be human led, limited and typically to be non-commercial in nature. Therefore, the use of AI training data which is for commercial use is palpably outside the fair dealing doctrine crafted by Indian **courts.**[18]

The Act also permits incidental electronic storage, such as temporary caching, where such storage is purely technical and ancillary to a **permitted use.**[19]These are inserted keeping in mind that, for routine digital process for internet functioning, sometime such facility is required. But AI training involves repeated storage of copyrighted works as datasets. Such training is not temporary or incidental and deliberate storing is essential to AI training process. In such a case the incidental storage exception cannot be used to justify the use of copyrighted works in AI training data. For this reason, within the purview of the Section 52 of the Copy Right Act and judicially crafted contours for application of fair use doctrine being implanted in Indian legal regime , it is evident that copyright law offers limited statutory support for AI training practices involving unauthorized copying of protected **works**[20].It may be reasonably argued that the AI processed data would not get any protective shield if generated in defiance to the stipulations of the Section 52 of the Indian Copyright Act.

Moral Rights, Performers' Rights and Identity in the Age of AI

Copyright law in India not only protects economic and creative interests of creators but also shields their creative integrity by invoking the doctrine of moral rights. The section 57 **of the Act**[21], empowers an author with the right to claim authorship and concurrently the right to prevent or claim damages for any distortion or modification of their works tarnishing honour and reputation of him inescapably bonded with his creations. This right exists independently of the ownership and survives even after the author has assigned the copyright. This underpins the law's emphasis on protection of integrity of personal attachment between an author and his creative expressions. As already argued , AI

training involves the ingestion of data and generation of outputs based on that data and these outputs become similar to the original works. While such outputs may not infringe on reproduction rights, they raise concerns regarding distortion and erosion of the author's identity and value of the original work. The Delhi High Court has upheld that the destruction or mutilation of an artistic work violates moral rights of its author emphasizing on the principle that creative works embody the **author's personality and honour.**[22]

Such non economic rights are further strengthened in the Copyright Act of 1957 after the amendment of the Act in 2012 which added "Performers' Rights" within its fold for legal protection. It has incorporated sections 38A and 38B to the Act which having protected performers' exclusive economic and moral rights on their **performances.**[23]This has laid down the principle that performances including voice and likeness are distinct contributions that deserve special protection. The Bombay High Court in **Arijit Singh**[24] case upheld that, the voice of a performer is an integral part of his identity and cannot be commercially used without his authorization. When AI is trained on recordings of a person's voice which can create fake versions of it, serious issues of consent, misuse, and breach of inalienable personality rights would arise. Even if this does not indicate copying of a copyrighted work in the usual way, it still misappropriates the performer's works and identity in a way, which is anti-thesis to principles of the law.

Mandates of performers' rights and moral rights transpire that, copyright law is concerned with creative works and performances, which are used in AI training without conforming to the procedural stipulations ignoring prior consent or showing apathy to integrity of expressions of the original authors . Such a situation indicates the violation of non economic rights of the authors, which is *sine qua non* for prevention of prospective or intrusivecavity on the reputation of an author. This means that , AI training cannot be judged only from the standpoint of law of copy right and fair dealing, but also for protection of personal dignity and reputation of an author as underpinned by the settled legal principles.

Intermediary Liability and Accountability in AI Training

Accountability in AI training is required to be examined along with the role of intermediaries for construction of percipient postulates for further scrutiny.As established under the Information Technology **Act (IT)** intermediaries receive a safe harbour as it shields them from liability of any third party for having posted on their platforms any offensive **contents**[25]. But compliance with the process under this statute for enjoying the privilege is required. This requirement is essential for ascertaining as to whether AI developers can avoid liability under copyright law in AI training. The Hon'ble Supreme **Court of India**[26] upheld that intermediaries are not required to monitor content actively. The liability for such actions would be enforceable only after the receipt of a Government notice or Court order in regard to any illegal activity having being taken place on their platforms. This stipulation is formulated to prevent censorship on expression of personal opinion for protecting free speech & expression and therefore, intermediary liability may be termed as qualified in the said context and the law is also settled to that extent.

AI developers alsoargue that they act as intermediaries for processing third party data and are therefore protected under provisions of Section 79 of the IT Act and should not be considered as primary infringers of copyright. They argue that unlike traditional intermediaries, the AI developers actively collect datasets for the sole purpose of training their AI models. A few scholars have observed that such collections are fully intentional and not just a mistake or byproduct, as it is the most essential part of AI **training**[27]**.** Therefore, this raises serious doubts on the contention of AI developers on claiming protection under Section 79 of the IT Act. Moreover, the protection under Section 79 is dependent on the compliance with conditions by an entity advocating for legal protection with the due diligence obligations laid down in the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. These rules explicitly state that intermediaries must act upon any illegal activities known to them and make sure that enough precautions are taken to prevent misuse. If AI developers are aware or made aware that copyrighted-work is being used in training of their AI models and if they continue to use the same, then they shall be no longer protected under Section 79. Therefore the logical argument should be that, protection granted under this Section is purely conditional and qualified and the law has not carved out any blanket exception for the same.

It may be reasonably inferred that the Section 79 may protect online platforms, but its application to AI developers is uncertain and judicially not delineated. The fact that AI training needs data collection is completely different from neutral intermediaries ,

where information or data is posted by third parties . This uncertainty therefore underpins the need for clearer accountability standards and hence specific legislation is needed to address this unique role played by AI developers involving such critical issues of Law.

Personal Data, Consent and AI Training under the Digital Personal Data Protection Act, 2023

AI training raises concerns over the use of personal data as well as scope of its protection under existing legal framework. In India, these concerns over personal data are governed by the Digital Personal Data Protection Act, 2023 (DPDP Act). It regulates the use of personal data in digital mode and the Act needs critical discussions in relation to personal data being processed without the consent of individuals in the AI Training modules. According to the DPDP Act, personal data can be processed or used if there is free and informed consent of the owner of the said **data.**[28] On the surface it seems that the Act is capable of regulating and restricting the AI developers from using personal data without the consent of owners, However the Act has significantly carved out exceptions , specifically, The Section **3(c)**[29] of the said Act excludes any such personal data that has been made public by the owner or is required to be publicly available from the protective shield and statutory protection is not available to the data principal (owner) in such an eventuality. Under such a situation a substantial portion of AI training data which is apparently personal data, is taken from publicly accessible sites such as social media, blogs, etc and therefore owners of data would not get protection under the said Act. Under such a situation use of the same would not amount to violation of the DPDP Act.

This creates a significant regulatory gap, even though DPDP Act is based on the consent centric approach. The exclusion of personal data that is publicly available may allow AI developers to use such data for AI training without any adverse legal implications. Individuals may make information public for limited purpose , but the Act has not impeded the use of such data for AI training, at least for leading the argument under that legal presumption. The Act provides limited grounds for processing personal data without consent, including for specified purposes such as compliance with law or **research, but it is often suppressed.**[30]The fact is that, these exceptions are narrowly framed and do not clearly accommodate commercial AI training. Unlike the European Union's GDPR (General Data Protection Regulatory), the DPDP Act does not recognize a general "legitimate interest" exception for processing/using data and also does not impose specific transparency or accountability obligations on AI developers with respect to training datasets. For this reason, while the DPDP Act marks an important development in Indian data protection law, it offers limited protection against the use of personal data in AI training. By excluding publicly available personal data from the protective framework and not specifying AI-specific safeguards, the Act leaves a substantial portion of AI training practices outside its regulatory reach. Such a visible gap has unearthed legal deficiencies as also observed under Indian copyright law and the need for a more coordinated legal response to AI training data in regard to maintaining integrity of personal digital data is truly required.

PERSPECTIVES ON AI TRAINING DATA: UNITED STATES AND EUROPEAN UNION-A COMPARATIVE STUDY

The comparative study of foreign laws with that of Indian Legal Regimes is an effective way of assessing India's legal position on AI training data *vis- a- vis* the issues of Copy Right Protection. The laws of USA and EU represent two completely different types of regulatory frameworks, which are used to understand and justify critical intersections of AI training and copyright protection globally. But there is a caveat as embodied in the unvarnished mandate of Law that, foreign legal doctrines cannot be replicated in the Indian legal ecosystem without rigorous scrutiny of facts and circumstances and situational dynamics incidental to the same.

A. United States of America: Fair Use and Judicial Flexibility

The USA follows a quite open ended fair use doctrine under the Copyright **Act.**[31] In contrast to India's purpose based approach, in USA, fair use is determined by applying the four factor balancing test. The test considers: purpose, nature of copyrighted work, amount invested and economic effects. Over the years courts in the US have interpreted fair use in a very expansive manner specially in cases involving technological **innovations**[32]. The AI developers in USA often rely upon the fair use doctrine to justify use of copyrighted work for training **AI**[33]**.** While the US courts are yet to decide on the legal standing of using copyrighted work for AI training under fair use exception , a few interim decisions of the courts indicate willingness of the courts to consider such data use/processing as a transformative process, if the output generated by such process does not replicate original works. The flexibility as observed

in the US Copy Right law is in -built with a purposive window, while Indian copyright law has not endorsed a broad or open ended fair use doctrine. The Indian courts have repeatedly declined to accept the argument for an expansive fairness doctrine and rather our legal regime inclines to endorse a strict statutory approach. Therefore, U.S. fair use jurisprudence provides little assistance to the Indian Legal System while evaluating the legality of AI training practices under the Indian Copyright Act, 1957.

B. European Union: Text and Data Mining

The EU has adopted a very structured approach and the EU Copyright Directives (DSM Directives) resulted in the introduction of Text and Data Mining (TDM) exceptions. The DSM allows the use of TDM for non commercial scientific **research**[34] and other pur**poses**[35] with the choice of opting out by copy right **holders**[36]. This statute attempts to maintain a balance between technological innovation and inalienable copy rights of authors. Moreover the EU's new AI **Act**[37] aims to introduce transparency and accountability for AI developers. India does not follow a similar approach and the Indian Copyright Act does not have any TDM **exception**[38]. In contrast to the EU, India relies on general copyright principles judicially upheld and statutorily delineated .Our legal regime is engineered accordingly for dealing with the issue , but visible gap still persists in theory and practice .

C. Implications for Indian Copyright Law

In comparison to the USA and EU laws, India has adopted a distinct approach to AI training of data without eclipsing the statutory delineations and interpretations of the fair use doctrine in regard to use of Copy Righted works by Courts of Law. In India, copyright law follows an approach with limited exceptions and protection of copy right is cemented for protection of intellectual creativity and commitment of the legal regime for upholding integrity of authors' intellectual creations within a structured pattern as being judicially interpreted and crystallized . But it would be fair to presume that , AI training cannot be legalized only through judicial interpretations with hyper technical calibration of facts unless Law specifically stipulates exceptions for the same in public interests and guiding principles are not pragmatically formulated .

Case Studies and Emerging Litigations on AI Training Data

Judicial approach as to the legality of the use of copyrighted data is still at a developing stage. But emerging legal disputes highlighting the approach of the judiciary on the use of AI training data have afforded us an opportunity to understand multifaceted issues involved on the subject.

A. ANI Media Pvt. Ltd. v. OpenAI[39]

The case of ANI media Private **Limited** has ushered in a new chapter in the particular domain as the first major litigation encompassing the issue of AI training data under copyright law. The ANI has accused the OpenAI of unlawfully storing, reproducing and using copy of its news content and data of the ANI company for training of their AI model known as ChatGPT. The case is still pending for final adjudication in Delhi High Court and the ANI News appealed to the High Court for an *ex parte* ad interim injunction for restraining the OpenAI from directly or indirectly using ANI Media's data. One of the major elements of this case is the concept of "machine unlearning", which means removing the influence of copyrighted works from the outputs generated by AI models. A few critical questions regarding implementation of the concept were raised in the court, even if it might be possible for unlearning copyrighted content or on probability of causing practical effects on the AI Modules, when such contents on which the AI is trained, **is unlearned.**[40] While OpenAI has effectively caused removal of AI media website from its system which was used as a source for training AI model, the aggrieved party (ANI Media) argued that the ChatGPT generated output is still reproducing the ANI media's copyrighted works under the pretext of having being available from third party sources. This case exposes a major structural mismatch between the copyright law and AI architecture. This case, therefore, implanted red flags for the courts to decide as to whether remedies based on **takedown principle and subsequent removal of offensive contents** would cater to requirements of the fair and reasonable adjudication on disputed issues.

B. Getty Images (US) Inc. v. Stability AI Ltd.[41]

In United Kingdom, Getty Images filed a case against Stability AI and they accused stability AI of using millions of copyrighted photographs to train the stable diffusion and image generation model without the consent of the company. The petitioner also stated that many of the AI generated outputs had Getty's watermark arguing that same could be clear evidence of unlawful copying. While some of the claims advanced by the Getty, in the initial stages were dismissed, the UK High Court has not yet delivered

final judgement on the same. This matter is scheduled for trial and it is expected that, the core issues would be unswervingly addressed for returning a reasonable decision by the Court. This case is of paramount significance, as the law point is expected to be constructed on the issue -- as to whether collection and storage of copyrighted works for AI training can be justified under current Copyright Law without consent/authorization .

C. Robert Kneschkev v. LAION e.V[42]

In this case, Hamburg Regional Court addressed the legal question of using copyrighted images for the creation of the LAION-5B dataset. This dataset is consisted of billions of image text pairs which were used for AI training. The plaintiff of this case accused the company of illegally downloading, scrutinizing and including his photographs in the data set. The court in this case, however, dismissed the claim of the plaintiff stating that the actions of the company were protected under the exceptions of the German Copyright Act, specifically text and data mining exception for scientific **research**[43]. Furthermore, the court also opined that the burden of proof falls upon the plaintiff in this case. It is the plaintiff who has to prove that, the exception does not apply to the instant case and the Court leaned in favour of the opponent by relying on the presumption to the said extent.

D. The New York Times Co. v. OpenAI and Microsoft[44]

The New York Times and various other news organizations have brought a litigation against OpenAI and Microsoft. The news organizations accused Microsoft and OpenAI of using copyright news content for training their AI models. The plaintiffs argued that copyrighted works were copied, stored and encoded during training and they also submitted that the generated outputs from those AI models had very closely resembled original copyrighted works. In the month of April 2025, the US District Court of New York rejected several motions brought by the defendant to dismiss the case and allowed the case to be proceeded with. In this case it was recognized that the training stage of AI involves storing and encoding of copyrighted works and the plaintiff is acquainted with the same as argued by the defendant. However, the court dismissed certain claims relating to unfair competition and abridgement, transpiring need of the heightened circumspection for extending liability for infringement of copyright and the case is yet to be decided and the final judgement is expected to indicate as to whether the fact in issue may be attributed **to the doctrine of fair use or no**t .

E. Studio Ghibli Style Creative Identity Controversy[45]

The case of studio Ghibli is a very well-known case throughout the world in the context of our discussions. It involved a series of AI generated images, replicating the aesthetic art style of the studio with the use of ChatGPT. Even though there was no judicial intervention against the same as yet, many critics have argued that effortless generation of Ghibli style images suggest that, AI models were trained on copyrighted Ghibli works without authorization. While copyright law, traditionally does not protect style, but this controversy highlights as to how the AI training could erode creative autonomy of the author of a work and can construct a duplicitous boundary between autonomous and mis-appropriated copyrighted works .

CONSENT, LICENSING AND POLICY MODELS

The legality of training data under AI modules may be construed in terms of our discussions as above and arguments posited are aimed at framing of the core issues and indicating probable remedies under an extended legal framework, where harmony, rather than acrimony, is expected to give leverage in a technology driven and knowledge propelled socio-economic ecosystem. We are also inclined to find out an efficacious remedy where fruits of intellectual labour and skills of a human being are not misappropriated by others transcending the legal trajectory intelligently embedded into it, the doctrine of fair use.

It has been established that storage and reproduction of copyrighted works for AI training does not fall under the exceptions given under Section 52 of the Indian **Copyright Act**[46]. Therefore, there is no doubt as to the essentiality of the consent prior to use of such works in AI training , but at the same time how the same is to be executed without impairing the technology driven ecosystem is also important. This paper will now delve into examining regulatory models capable of addressing AI training without affecting the mandate of the Copyright law.

I. Consent as the Structural Foundation

The copyright law of India is based on consent of the author and the law vests exclusive rights on authors and owners of copy righted works. Use of any copyrighted work will be legal, if authorized by the right holder or permitted by statute under a particular

situation.[47] Unlike other legal regimes, Indian law has implanted prior consent of the author as mandatory condition before using his copyrighted creations for AI Training so far as the current statutory framework is concerned . However, the current AI training process operates outside this structure. Copyrighted works are collected, stored and used without authorization. If AI training is to be legalized, it should be brought within a legal framework which will ensure control of an author over his copyrighted work instead of marginalization of it.

II. The Impracticality of Individual Authorization

While consent involves a legal process of licensing by an individual author for use of his copyrighted works, the execution of it appears to be quite difficult in the context of AI training. For AI developers prior negotiation with each right holder would be challenging because training is based on millions of works collected from various platforms and jurisdictions. But at the same time, the AI data collection process unless regulated by statutory process, will continue to shake the foundation of copyright law and that situation would be antithesis to harmonious co-existence of two regimes.

III. Option in Licensing Models

Option in licensing is one of the most practical approaches, as argued by some scholars. Under this approach, authors of works would voluntarily license their works for AI training. Such licenses shall be administered by Right Management **Organisations**[48]. Incorporation of the scope for according option in models also aligns with consent principles, of course with structural limitations. As large AI developers shall dominate the majority of the datasets, there is possibility that smaller entities may have a hard time while executing this option and a fair and reasonable solution may be difficult to be achieved .

IV. Collective Licensing as a Structurally Compatible Solution

Many copyright societies or organizations usually license music, broadcasting and performance rights which may . balance access with **compensation**[49]**.** If applied to AI training, such an approach may assist AI developers in accessing large datasets through standard licensing. While the authors of creative works are compensated fairly, such an approach may reduce transaction cost, provide legal conformity and uphold central role of consent in accessing copy righted works.

V. Statutory Licensing and Legislative Intervention

In situations where obtaining permission from every Right holder is unrealistic, legislatures may consider introducing a narrowly defined statutory license for AI training. The Indian copyright law has already mandated statutory and compulsory licence in areas where individual negotiation is impractical but public interest requires only a **regulated and limited access.**[50] Under such a framework, AI developers may be allowed to use copyrighted works for training in return for mandatory payment, maintaining transparency about data use and compliance with prescribed safeguards. However, such a license should be carefully put in to practice and for limited purpose only. If framed too broadly, it will risk the protective control of an author over his work(s) and core principles of copyright law would be infructuous to that extent .

VI. Smart Contracts for licensing and royalty distribution

Collective and Statutory licensing for AI can be supported by smart contracts. Smarts contracts have the capability of automatically calculating royalties and applying terms which makes large scale datasets more efficient, provided that required consent is **obtained**[51]**.** Smart contracts do not create legal rights on their own; those are used to implement obligations that exist in Copyright Law . This approach can be used to handle high volumes of datasets.

CHALLENGES IN ENFORCEMENT AND REGULATION

Even though the legal position of AI training data is clear, the enforcement of copyright obligations is difficult. Such difficulties do not arise from law but from technical design of AI. There is lack of transparency and integrity in regard to cross border nature of the data being used in AI Training. The AI training usually takes place in complete secrecy where datasets are rarely disclosed. This eliminates the chance to resist such use by the Copy Right holders . It is observed that , due to this process, right holders face a hard time in proving when and how their copyrighted works have been used before the Court of Law, as and when necessary . Obviously it is felt that, by placing the burden of leading evidence on authors/creators of copy righted works for sustaining allegation of the infringement of Copy

Right , the authors on many occasions become helpless and technicalities of legal process widen the gulf in unearthing actual facts and compliance of nitty-gritty of evidence for adjudication in the Court of Law. As a result of insightful analysis of legal principles during the instant academic discussions , there is no difficulty in assuming that traditional enforcement rules can only unearth limits of output based acceptance of generative AI functioning. But where AI system becomes intrusive to right holders , scope of instant remedy becomes uncertain . Generative AI systems are designed in a way, in which they do not produce an output that is similar to the copyrighted work. Even if unauthorized copying occurred during training, the law only focuses on the nature and types of outputs. As a result the copyright infringement that took place during the training, remains unnoticed, exposing the limitations of meaningful solutions for otherwise visible acts of copying and distribution. These problems are further multiplied by jurisdictional complexities as AI developers' work in multiple countries, which makes enforcement by Indian Right holders against foreign entities quite difficult. It is clear that traditional enforcement rules are not intelligently framed for addressing such issues.

Therefore the proposed technical solutions can be dataset deletion or machine unlearning which would offer limited relief to the right holders, if issue of licensing or consent is not addressed legally and immediately. Once a model has been trained, removal of influence of copyrighted work from AI becomes technically uncertain. At the same time, it becomes difficult to supervise proper control over information in regard to training of the AI systems and the authors' indifference to assertion of their copy rights also stultify the scope of processing legal remedy. Existing limitations show that, only legal rules are not enough for effective regulation of AI training , which requires targeted legal and institutional reforms for affording remedy against opacity and asymmetry in AI outputs in regard to use of copyrighted contents in the process.

RECOMMENDATION FOR EFFECTIVE MECHANISM IN INDIAN CONTEXT

From the discussions, it may be safely presumed that, the difficulty in regulating AI training data lies not in conceptual uncertainty, but in structural cloudiness. There is evident imbalance and mismatch between traditional copyright remedies and application of settled legal principles for affording effective remedies for infringement of the copyrighted contents and personal data by automated AI systems under new operational dynamics. The Indian copyright law would require to incorporate procedural and structural changes for addressing some critical issues, primarily as to how AI system is trained and developed. There is requirement of two tier of reforms for finding out meaningful solutions, one being the core structural reforms which is conventional in nature and the second, may be with unconventional regulatory mechanism. The following suggestions may be indicative relying on the discussions with reference to authoritative sources already mentioned-for finding out some effective solutions for meaningful co- existence of Copy Right law and AI training :

I. Conventional Reforms

A. Statutory Recognition of AI Training

The Copyright Act 1957 should be amended to acknowledge reproduction and electronic storage facilitated for AI training as legally acceptable exception. The existing legal framework is not enough to respond AI training and only explicit regulations would remove uncertainties triggered by intersections in legal principles and procedures . It may be reasonably presumed that AI training with statutory provisos within the framework of copyright may be accommodated by accepting consent based preconditions.

B. Transparency obligations

Transparency is essential for meaningful consent and enforcement of Copy Right Law. AI developers should be brought under a statutory network to reveal or disclose information regarding the AI training data being used in the process. However, this should be stipulated that such disclosures shall not unearth any trade secrets of the company but would be indicative enough to allow copyright holders to recognize and assert affirmative actions, if their works were used without authority.

C. Harmonizing Copyright law and DPDP Act, 2023

Copyright law should work together with the DPDP Act, 2023 and publicly available data may be admissible for AI training relying on facts and circumstances. Functional coordination between two wings of law may plug off loopholes in regard to enforcement of Copy Right and protection of personal data as harmonious co- existence of copy rights and protection of personal data should be reflected in a statutory framework for fair use in AI platforms.

II. Unconventional Reforms

A. Right to Audit AI Training Processes

Copyright law could not introduce a system where the creators of the copyrighted works are allowed to access the training process of the AI, where they would get access to the data that is being used to train the AI. This system can be introduced as a procedural safeguard for the copyright holder. Such a right holder would not gain access to full data source, but would require the AI developer to verify works that have already been used in the AI training and also to verify the sources and timeframe.

B. Presumption of Use for High Probability Training Sources

When models are trained on datasets which are available on large public platforms a rebuttable presumption may be constructed that works hosted on such platforms and used in training may be admissible, unless evidence is led by the aggrieved party to the contrary. Under such a situation, the AI developers would be under an obligation for maintaining integrity of the copied contents as well as moral rights of creators. The rebuttable presumption, if incorporated in to the legal framework, the burden of proof will require to be discharged by the parties reasonably and that would also insulate the concerns of creators of the copyrighted works to a reasonable extent.

C. Human Creative Dependency Index

Legal mechanism may be introduced which would help the court to assess the level or degree of dependency of an AI system on copyrighted works based on multiple factors such as data set composition, reliance on copyrighted works and potential for replacement. The legal regime can also measure the degree of dependency on copyrighted works and also in turn safeguard genuine AI tools from over regulation.

D. Time bound Training Immunity

Another approach that India can adopt is -- time bound immunity, rather than giving permanent exceptions. Introduction of time-bound training with immunities for experimental models may be subjected to automatic expiry after a particular interval. Immunity would expire if model is commercialized for profit and full compliance obligations would be mandatory for the AI developers. In this way, innovation would be encouraged without permanently divesting the authors of control over copyrighted contents coupled with recognized neighboring rights and privileges and also may be effective for avoiding irreversible effects of exceptions on some situations.

E. Cultural Impact Assessments for Generative AI Models

Before the release of AI models into the public domain, there should be a stipulation for the AI developers for conduct of a rigorous evaluation on issues of their models affecting copy right of artists, prospects of particular languages and cultural communities. The cultural impact assessment is similar to environmental impact study to some extent . Furthermore, this would protect cultural interest as a critical component of public interest warranting protection. Such an approach would shift AI regulation beyond commercial interests and competitions by necessitating the regulators to safeguard erosion of cultural values and marginalization of creative efforts and skills of authors of copy righted works.

CONCLUSION

This paper has examined the legality of AI training data under Indian copyright law by focusing on relevant issues pertaining to ownership, consent, and accountability. We have examined whether copyrighted works used by the AI developers for AI training amount to copyright infringement and how personal data used in AI training may claim exemptions under the Digital Personal Data Protection Act, 2023 along with Rules framed for the implementation of the Act. Further, in the paper efforts are also made to compare India's response to AI training data with that of United States of America and the European Union. Our scrutiny of relevant Laws and facts have transpired that the AI training necessarily facilitates the reproduction and electronic storage of protected creative works of authors and under the copyright law of India, the same means infringement of legal rights of the copyright holders, irrespective of whether the output generated by AI models is similar to such works or not. In summary, the foregoing analysis has clearly established that , the AI training involving copyrighted works should mean **infringemen**t within the meaning of Sections 13, 14 and 51 of the Copyright Act, 1957. It is evident that the limited and legally settled fair dealing exceptions under Section 52 of the Copy Right Act will not include large-scale commercial AI training practices. The Digital Personal Data Protection Act, 2023 by excluding publicly available personal data from its protective shield , has implanted regulatory gap in the context of AI training on personal data **vis – a- vis** statutory obligations of

the law of copy right including moral rights integral to authorship. Our discussions also corroborate that, the Section 52 of the Indian Copyright Act 1957 extends only limited protection as regard to use of copyrighted works without consent for non commercial purpose and in public interest. Concurrently we have also supplemented that, the Indian legal system follows a specific fair dealing framework with the **read down alignment, banking on judicially calibrated principles**. The emerging case laws including Getty Images vs. Stability AI and ANI vs. Open AI *inter alia* indicate that , courts are required to adjudicate on issues encompassing the structural mismatch between obligations of Copyright Law and subtleties of AI architecture for prevention of arbitrariness in functioning and governance of AI Regime. Considering this important aspect , it may be reasonably presumed that Indian Law has already provided a strong foundation to underpin the argument for regulating AI training data without substantially affecting the legal and statutory rights of the Copyright holders. But the gap in regard to fair and harmonious enforcement of legal principles in the AI data training structure has not been effectively bridged till date. The instant academic analysis and interpretations have identified such issues rationally and for meaningful co- existence of Copyrighted works of authors and autonomy of AI modules, our suggestions may safeguard efficacious solutions.

We, therefore, envisage systematic and innovative reforms in the legal regime for greasing the joints of copyright compulsions and obligations of a responsive AI regime for accelerating the functioning of an accountable AI structure in the coming days, where creative integrity is supplemented by technological progress without eclipsing accepted silhouettes of Intellectual Property Rights, of which copyright is a pivotal branch.

As upshot of discussions, it may be concluded that AI Regulation is urgently required focusing on collection, storage and *bonafide* use of copyrighted contents without impairing basic principles of Law of Copy Right and protection of personal data . An Indian specific approach is necessary by introducing reforms in the copyright regime where technological innovations may be promoted and also creative autonomy of an author is not derailed by implanting consent based preconditions in AI data training exercise on processing copy righted information.

REFERENCES

1. Copyright Act 1957, s 17.
2. Copyright Act 1957, ss 13–14, 52.
3. Copyright Act 1957, s 14.
4. Eastern Book Company v. DB Modak (2008) 1 SCC 1; David W Opderbeck, 'Copyright in AI Training Data: A Human-Centered Approach' (2024) 76(4) Oklahoma Law Review 951–1010; Tim W Dornis, 'The Training of Generative AI Is Not Text and Data Mining' (forthcoming 2024) European Intellectual Property Review (English translation, available at SSRN) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4993782 (accessed on 27 February 2026).
5. Pamela Samuelson, 'Does Using In-Copyright Works as Training Data Infringe?' (2025) 68(11) Communications of the ACM 28–30; Mark A Lemley and Bryan Casey, 'Fair Learning' (2021) 99 Texas Law Review 743; Andrew W Torrance and Bill Tomlinson, Training Is Everything: Artificial Intelligence, Copyright, and Fair Training (2023).
6. Eastern Book Company v DB Modak (2008) 1 SCC 1.
7. Copyright Act 1957, s 13.
8. Eastern Book Company v DB Modak (2008) 1 SCC 1.
9. David W Opderbeck, 'Copyright in AI Training Data: A Human-Centered Approach' (2024) 76(4) Oklahoma Law Review 951–1010; Li You and Han Luo, 'Copyright Implications and Legal Responses to AI Training: A Chinese Perspective' (2025) 14 Laws 43.
10. Copyright Act 1957, s 2(d).
11. Copyright Act 1957, s 57.
12. Amarnath Sehgal v. Union of India, 2005 (30) PTC 253 (Del).
13. Copyright Act 1957, s 13; Shridul Gupta, 'Fair-Use Doctrine: Copyright Challenges Posed by AI Generative Technology' (2025) 14 Christ University Law Journal.
14. Copyright Act 1957, s 14.
15. Copyright Act 1957, s 51.
16. Copyright Act 1957, s 52; Weiyi Xia, 'Fair Use of Training Data in Generative Artificial Intelligence' (2025).
17. Super Cassettes Industries Ltd v Chintamani Rao, 2010 SCC OnLine Del 2736
18. Wiley Eastern Ltd v Indian Institute of Management, (1996) 1 SCC 31; Kailash Chauhan, 'Generative AI, Text & Data Mining and the Fair Dealing Doctrine' (2025) 30 JIPR 77.
19. Copyright Act 1957, s 52(1)(b)-(c).
20. Pamela Samuelson, 'Does Using In-Copyright

Works as Training Data Infringe?' (2025) 68(11) Communications of the ACM 28; Andrew W Torrance and Bill Tomlinson, Training Is Everything: Artificial Intelligence, Copyright, and Fair Training (2023).
21. Copyright Act 1957, s 57.
22. Amarnath Sehgal v Union of India 2005 (30) PTC 253 (Del).
23. Copyright Act 1957, ss 38A and 38B
24. Arijit Singh v. Codisto Ventures LLP, COM IPR SUIT(L) No. 23443 of 2024; Bedanta De, 'Training AI Models on Copyrighted and Personal Data: Reconciling Fair Use and Privacy Rights' (2025) VII(III) Indian Journal of Law and Legal Research 8869.
25. Information Technology Act, 2000, s 79.
26. Shreya Singhal v. Union of India, (2015) 5 SCC 1
27. Adam Buick, 'Copyright and AI Training Data—Transparency to the Rescue?' (2025) 20 Journal of Intellectual Property Law & Practice 182; S Alex Yang and Angela Huyue Zhang, 'Generative AI and Copyright: A Dynamic Perspective' (arXiv, 2024) https://arxiv.org/abs/2402.17801 (accessed on 21 April 2026).
28. Digital Personal Data Protection Act, 2023, ss 6 & 2(t).
29. Digital Personal Data Protection Act, 2023, s 3(c).
30. Digital Personal Data Protection Act, 2023, s 7; Shivi Kanjolia, 'Transforming Innovation into Fairness: Tackling Legal Challenges in AI-Enhanced Education in India' (2025) Indian Journal of Integrated Research in Law.
31. Copyright Act 1976 17 (US), 17 USC § 107.
32. Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015)
33. Mark A Lemley and Bryan Casey, 'Fair Learning' (2021) 99 Texas Law Review 743.
34. Directive (EU) 2019/790 on Copyright and Related Rights in the Digital Single Market, art 3.
35. Directive (EU) 2019/790 on Copyright and Related Rights in the Digital Single Market, art 4.
36. Eleonora Rosati, 'Is Text and Data Mining Synonymous with AI Training?' (2024) 19 Journal of Intellectual Property Law & Practice 851; Tim W Dornis, 'The Training of Generative AI Is Not Text and Data Mining' (2024).
37. Regulation (EU) 2024/1689 (Artificial Intelligence Act).
38. Tim W Dornis, 'The Training of Generative AI Is Not Text and Data Mining' (forthcoming 2024) European Intellectual Property Review (English translation, available at SSRN)https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4993782(accessed 27 February 2026).
39. Bharatwaj Ramakrishnan, 'ANI v OpenAI: Time to Talk About "Machine Unlearning"' (SpicyIP, 2024) https://spicyip.com/2025/01/ani-v-open-ai-time-to-talk-about-machine-unlearning (accessed 27 February 2026).
40. ANI Media Pvt Ltd v. Open AI Inc., **CS(COMM) 1028/2024 (Del. HC)**
41. Getty Images (US), Inc. v. Stability AI Ltd., [2023] EWHC 2863 (Ch) (pending)
42. Kneschke v. LAION e.V., Case No. 310 O 227/23 (Landgericht Hamburg Sept. 27, 2024) (Ger.).
43. Urheberrechtsgesetz (Germany) s 60d.
44. The New York Times Co v OpenAI LLC No 1:23-cv-11195 (SDNY, 4 April 2025).
45. Arya Gerard, 'Totoro in Trouble? Assessing the Legal Implications of AI-Generated Ghibli Images' (Law Gazette, 2025) https://lawgazette.com.sg/feature/totoro-in-trouble(accessed 27 February 2026).
46. Copyright Act 1957, s 52.
47. Copyright Act 1957, ss 13-14.
48. Li You and Han Luo, 'Copyright Implications and Legal Responses to AI Training: A Chinese Perspective' (2025) 14 Laws 43.
49. Copyright Act 1957, ss 33-36.
50. Copyright Act 1957, ss 31 & 31D.
51. Geraldine O Mbah, 'Smart Contracts, Artificial Intelligence and Intellectual Property: Transforming Licensing Agreements in the Tech Industry' (2024) 5 International Journal of Research Publication and Reviews 317; Anirban Mukherjee and Hannah Hanwen Chang, 'Copyright Laundering Through the AI Ouroboros' (arXiv, 2026) https://arxiv.org/abs/2601.02631 (accessed 27 February 2026).

TABLE OF LEGISLATIONS

1. Copyright Act 1957 (India).
2. Copyright Act 1976, 17 USC § 107 (United States). Digital Personal Data Protection Act 2023 (India).
3. Directive (EU) 2019/790 on Copyright and Related Rights in the Digital Single Market.
4. Information Technology Act 2000 (India).
5. Regulation (EU) 2024/1689 (Artificial Intelligence Act). Urheberrechtsgesetz (UrhG) (Germany).

TABLE OF CASES

1. Amarnath Sehgal v Union of India 2005

(30) PTC 253 (Del).
2. ANI Media Pvt Ltd v OpenAI LLC CS(COMM) 1028/2024 (Del HC) (pending).
3. Arijit Singh v Codisto Ventures LLP COM IPR SUIT (L) No 23443 of 2024 (Bom).
4. Authors Guild v Google Inc 804 F 3d 202 (2d Cir 2015).
5. Eastern Book Company v DB Modak (2008) 1 SCC 1.
6. Getty Images (US) Inc v Stability AI Ltd [2023] EWHC 2863 (Ch).
7. Robert Kneschke v LAION eV Case No 310 O 227/23 (Landgericht Hamburg, 27 September 2024).
8. Shreya Singhal v Union of India (2015) 5 SCC 1.
9. Super Cassettes Industries Ltd v Chintamani Rao 2010 SCC OnLine Del 2736.
10. The New York Times Co v OpenAI LLC No 1:23-cv-11195 (SDNY, 4 April 2025).
11. Wiley Eastern Ltd v Indian Institute of Management (1996) 1 SCC 31.