COPYRIGHT INFRINGEMENT IN AI TRAINING: PROTECTING CREATORS IN THE DIGITAL ERA

G J Annie Beulet Mary, Assistant Professor, GTN Law College, Dindigul

ABSTRACT

The rapid development of generative artificial intelligence has transformed the creation, processing and dissemination of digital content, while simultaneously creating complex challenges for copyright law. Contemporary AI systems are trained using extensive datasets that may contain books, articles, photographs, artistic works, music, films, software and other copyright-protected materials. The collection, reproduction, processing and storage of such works during AI training raise fundamental questions concerning authorization, fair dealing, infringement, ownership, attribution and compensation of creators. In India, the Copyright Act, 1957 does not contain a specific provision governing the use of copyrighted works for training generative AI models, creating considerable uncertainty regarding the application of existing copyright principles to emerging technologies. This study examines the legal implications of using copyrighted works in AI training and evaluates whether the existing Indian copyright framework provides adequate protection to creators in the digital era. It analyses the exclusive rights of copyright owners, the scope of statutory exceptions, particularly fair dealing and potential liability of AI developers and technology companies. The study further examines emerging judicial and regulatory approaches in the United States and European Union to identify comparative solutions for India. It argues that technological innovation and copyright protection should not be treated as competing objectives. A balanced framework incorporating transparency in training datasets, appropriate licensing mechanisms, creator compensation, rights-reservation or opt-out mechanisms and accountability of AI developers is necessary. The legal reforms capable of protecting creative interests while supporting responsible development and deployment of artificial intelligence.

Keywords: Artificial Intelligence; Copyright Infringement; AI Training Data; Creator Protection; Generative AI.

INTRODUCTION

The rapid emergence of generative Artificial Intelligence (AI) has transformed the manner in which digital content is created, processed, reproduced and distributed, creating significant challenges for contemporary copyright law. Modern AI systems depend upon extensive datasets to learn patterns, language structures, visual characteristics, musical elements and other forms of expression. These datasets may contain substantial quantities of copyright-protected books, journal articles, photographs, paintings, music, films, software and online publications, often raising questions about whether such materials can be copied and processed for AI training without the authorization of copyright owners. The legal difficulty becomes particularly significant because AI training may involve multiple stages of downloading, copying, storing, preprocessing and computationally analysing protected works, while the resulting AI system may subsequently generate outputs that resemble existing creative works. In India, the Copyright Act, 1957 provides exclusive rights to copyright owners and recognizes certain statutory exceptions, including fair dealing under Section 52[1]; however, it does not expressly establish a comprehensive framework governing the use of copyrighted works for training generative AI models. This creates uncertainty regarding the legality of large-scale commercial AI training and the potential responsibility of technology developers. The issue also raises broader concerns regarding transparency of training datasets, attribution, consent, licensing, remuneration and the ability of individual creators to control the digital use of their works. Internationally, jurisdictions such as the United States and the European Union are developing different approaches to address the tension between technological innovation and copyright protection. Against this background, the present study examines the legal implications of using copyright-protected works for AI training, analyses the adequacy of the existing Indian copyright framework and evaluates emerging international approaches. It further explores possible mechanisms for protecting creators while preserving technological innovation and proposes a balanced regulatory approach based on transparency, licensing, accountability and fair compensation.

CONCEPTUAL FRAMEWORK OF AI TRAINING AND COPYRIGHT

The relationship between artificial intelligence and copyright begins with the manner in which AI systems acquire and process information. Generative AI models depend on large datasets

[1] Section 52 of Copyright Act, 1957 https://share.google/v4XysoVCyCmohlPnh

containing text, images, music, software and other forms of human-created expression. A substantial portion of such material may be copyright-protected, particularly where datasets are compiled through web scraping or other large-scale collection methods. WIPO notes that publicly available online sources can contain significant quantities of copyright-protected material used in AI training, creating unresolved questions concerning infringement, licensing, attribution and compensation. In the Indian context, these questions must be examined against the Copyright Act, 1957, particularly the exclusive rights of copyright owners and the statutory exceptions contained in Section 52.

2.1 Meaning and Nature of Generative AI

Generative Artificial Intelligence refers to AI systems capable of producing new content, including text, images, audio, video, software code and other expressive material in response to user instructions. Unlike conventional software that operates through predetermined rules, generative AI generally identifies patterns from extensive datasets and uses learned statistical relationships to generate outputs. Large language models, image-generation systems and multimodal models are prominent examples. The legal significance of generative AI lies in its dependence upon pre-existing human-created material. The quality and diversity of training data substantially influence the capabilities of an AI model. Consequently, copyright law becomes relevant before the AI-generated output is even produced. The central question is whether the lawful acquisition and computational processing of protected works during training requires authorization from copyright owners. WIPO has recognized this as an area of continuing legal uncertainty across jurisdictions.

2.2 AI Training and Training Datasets

AI training is the process through which a model analyses large quantities of data to identify patterns, relationships and structures that enable it to perform particular tasks. A training dataset may consist of text, books, news articles, photographs, artwork, recordings, software code and other digital material. Data may be obtained from licensed databases, public-domain sources, user-generated repositories or publicly accessible websites. The fact that material is available online, however, does not automatically remove copyright protection. During dataset preparation, works may be downloaded, copied, converted, filtered, annotated or stored. These activities can therefore intersect with copyright owners' reproduction rights. The legal analysis depends upon the nature of the work, manner of acquisition, purpose of use, applicable

statutory exceptions and whether authorization or a licence exists.

2.3 Use of Copyrighted Works in AI Training

The use of copyrighted works in AI training represents the central legal controversy. AI developers may reproduce protected material while collecting and preparing datasets and subsequently process that material during model training. The important distinction is between **access to a work** and **lawful authorization to reproduce or exploit it**. Public accessibility does not necessarily mean that copyright restrictions disappear.

Under the Indian Copyright Act, 1957, Section 14[2] provides copyright owners with exclusive rights over specified acts concerning protected works, while Section 51 identifies circumstances in which copyright is infringed. Section 52, however, creates specified exceptions, including fair dealing for particular purposes. The provision expressly recognizes fair dealing for purposes such as private or personal use including research, criticism or review, subject to the statutory conditions.

The difficulty is that **commercial generative-AI training does not neatly correspond to the purposes expressly identified in Section 52**. Therefore, whether a particular AI-training activity qualifies as an exception cannot simply be assumed. Factors such as the purpose and character of the use, the nature of the protected work, the extent of copying and the consequences for the copyright owner's legitimate interests require careful examination.

2.4 Copyright Interests of Creators

Copyright protects the economic and, in appropriate circumstances, personal interests of creators by granting legally enforceable rights over their original works. Creators of literary, artistic, musical and other protected works may have legitimate interests in controlling reproduction, adaptation, communication and commercial exploitation of their creations.

AI training creates a new dimension to these interests because creators may have little knowledge about whether their works have been incorporated into training datasets. WIPO has highlighted growing concerns regarding rights management, attribution and compensation as

[2] Copyright Act, 1957, Section 14, 51 and 52 https://share.google/xYLjTS64ONmgGztRe

copyright-protected works become increasingly important to AI development.

The issue is therefore not limited to proving traditional copying. It also concerns **control over digital exploitation**. A creator may argue that the commercial use of their work to develop a competing technological product should require authorization or compensation, particularly where the AI system generates commercially valuable outputs. At the same time, excessively broad restrictions could impede legitimate research, innovation and development of beneficial AI technologies. Copyright law must therefore maintain an appropriate equilibrium between creator protection and technological progress.

2.5 Relationship between AI Training and Reproduction of Copyrighted Works

The relationship between AI training and reproduction is particularly significant because training commonly involves the creation of copies at different stages. A copyrighted work may be downloaded from an online source, transferred into a dataset, stored on servers, transformed into a machine-readable format and processed repeatedly during training. Each stage may raise different legal questions concerning reproduction and applicable exceptions.

The Indian Copyright Act therefore provides the starting point for assessing whether such acts fall within the copyright owner's exclusive rights or a statutory exception. Section 52 contains specific circumstances in which certain acts do not constitute infringement, but it does not expressly establish a general exception for commercial generative-AI training.

Importantly, **training-stage reproduction and output-stage infringement should be analysed separately**. A model may be trained on protected works without necessarily reproducing an identifiable work in its output; conversely, an AI system may generate material that is substantially similar to a protected work, creating a separate infringement question. WIPO similarly recognizes that infringement concerns may arise from the scraping and use of works for training, the trained models themselves and their outputs, with the precise legal position varying between jurisdictions.

Accordingly, AI training cannot be treated as automatically lawful merely because the ultimate output is newly generated. The legality of the underlying acquisition, reproduction and processing of copyrighted works must be independently evaluated under the applicable copyright framework. This distinction provides the foundation for analysing the **Indian legal**

position, potential infringement, available defences and the need for creator-protection mechanisms.

COPYRIGHT FRAMEWORK GOVERNING AI TRAINING IN INDIA

The legal regulation of AI training in India currently depends primarily on the existing copyright framework rather than on an AI-specific copyright statute. The **Copyright Act, 1957** grants copyright owners exclusive rights over protected works and provides limited statutory exceptions to infringement.

This creates an important legal question: *when AI developers collect, copy, store and process copyrighted works as training material, do those activities fall within the copyright owner's exclusive rights or can they be justified under an existing exception?* The issue becomes particularly significant because Section 52 does not expressly establish a broad commercial exception for AI training. Therefore, the legality of AI training must presently be assessed by applying existing principles to the technological process.

3.1 Copyright Protection under the Copyright Act, 1957

The **Copyright Act, 1957** is the principal legislation governing copyright protection in India. Copyright subsists in original literary, dramatic, musical and artistic works, cinematograph films and sound recordings, subject to the statutory requirements of the Act. The framework is technologically neutral in several respects and therefore applies to works existing in digital as well as conventional forms. Section 14 defines copyright as the exclusive right, subject to the Act, to undertake or authorize specified acts in relation to a work or any substantial part of it. For literary, dramatic and musical works, these include reproduction, issuing copies, communication to the public, translation and adaptation. Importantly for AI, Section 14 expressly recognizes **storing a work in any medium by electronic means** as part of the reproduction right. This is significant because AI training may involve downloading digital works, converting them into machine-readable formats, storing them in datasets and repeatedly processing them. Consequently, the fact that a work is available online does not automatically mean that it can be freely copied for commercial AI development.

3.2 Exclusive Rights of Copyright Owners

The exclusive rights granted under Section 14 form the foundation for determining potential

infringement arising from AI training. Copyright owners generally have the right to control the reproduction, distribution, communication, translation and adaptation of their protected works, depending on the category of work. Section 51 provides that copyright is infringed, among other circumstances, when a person without a licence or authorization undertakes an act that falls within the exclusive rights granted to the copyright owner.

For AI training, **reproduction is particularly important**. A developer may copy thousands or millions of works into a dataset even though the final AI model does not reproduce those works verbatim. The legal question is therefore not restricted to the final output; the initial copying and subsequent processing of protected material may independently require legal justification.

At the same time, infringement cannot be presumed merely because copyrighted material exists somewhere within a dataset. The specific work, nature of copying, purpose of use, authorization, applicable exception and other relevant circumstances must be examined. This makes AI-training disputes substantially more complex than conventional instances of unauthorized copying.

3.3 Reproduction and Adaptation Rights

The reproduction right has particular relevance because Section 14(a)(i) expressly covers reproduction of literary, dramatic and musical works in any material form, including **electronic storage**. Section 14 also recognizes adaptation rights in relation to relevant categories of protected works. AI training can involve several forms of reproduction. A copyrighted article may first be downloaded from a website, copied into a dataset, converted into another format and stored on servers. These activities may potentially engage the reproduction right even though the ultimate purpose is computational learning rather than human consumption of the work. The question of **adaptation** is more complicated. Merely analysing a work to identify statistical patterns does not necessarily mean that the work has been legally adapted. However, where an AI system produces an output substantially based upon protected expression, questions of reproduction or adaptation may arise separately.

3.4 Fair Dealing under Section 52

Section 52 provides important statutory exceptions to copyright infringement. Under Section 52(1)(a), fair dealing with a work, other than a computer programme, is permitted for specified

purposes including **private or personal use including research, criticism or review and reporting of current events and current affairs**, subject to the statutory conditions. The provision also addresses certain forms of electronic storage connected with those purposes.

This provision creates a significant unresolved question for AI training. A research institution using a limited dataset for non-commercial research may present a different legal situation from a commercial technology company training a large generative-AI model using millions of copyrighted works to develop a profitable service.

The present wording of Section 52 does **not expressly mention artificial intelligence, machine learning, text and data mining or commercial AI training**. Therefore, it would be difficult to automatically treat all AI training as fair dealing.

The purpose of the use becomes particularly important. AI development may have scientific or research dimensions, but commercial deployment and monetization could make the legal analysis more difficult. The absence of an express AI-training exception consequently creates uncertainty for both creators and technology developers.

3.5 Computer Programs, Databases and Digitally Stored Works

AI systems themselves frequently involve computer programs, while their training datasets may contain databases and large collections of digital works. The Copyright Act specifically recognizes **computer programmes as literary works**. The statutory definition of computer programme covers a set of instructions expressed in words, codes, schemes or other forms capable of causing a computer to perform a task or achieve a particular result.

The Act also provides special exceptions concerning computer programmes under Section 52(1)(aa)–(ad). These provisions permit certain acts by lawful possessors, including specified copying or adaptation necessary for using a programme, backup purposes, interoperability and examination of its functioning.

However, these exceptions should not be interpreted as a general authorization to use any computer program for commercial AI training. Their language is directed toward particular purposes and circumstances.

Databases present another important issue. A dataset may contain individual copyright-

protected works, structured compilations, software or other protected material. Consequently, AI developers must consider not merely whether information is publicly accessible but also whether the **underlying material or compilation enjoys copyright protection** and whether copying it falls within an applicable statutory exception.

3.6 Applicability of Existing Copyright Law to AI Training

The existing Indian copyright framework can therefore be applied to AI training, but it leaves several important questions unresolved. The Act provides strong protection for reproduction and related rights, while Section 52 establishes limited exceptions; neither provision was specifically designed for large-scale machine-learning systems.

Three stages require separate consideration:

1. **Data acquisition:** whether collecting and downloading copyrighted works without authorization constitutes an unauthorized reproduction.

2. **Model training:** whether copies created, stored or processed during computational training are protected by an exception or require authorization.

3. **AI output:** whether an output reproduces a substantial part of a protected work or otherwise infringes copyright independently of the legality of the training process.

This distinction is essential because **lawful access to a work does not necessarily confer an unrestricted right to reproduce it for commercial AI training**. Conversely, the mere presence of copyrighted material in a training dataset should not automatically establish infringement without examining the specific statutory rights and exceptions involved.

The principal limitation of the current framework is therefore **legal uncertainty rather than complete absence of copyright protection**. Indian copyright law already contains mechanisms capable of addressing unauthorized reproduction, but it does not clearly determine how those mechanisms should operate when billions of digital works are processed by AI systems. A future framework may therefore need to address **training-data transparency, licensing, creator consent, text-and-data-mining exceptions, opt-out mechanisms and remuneration models**, while preserving legitimate research and technological innovation.

For the purposes of this research, the most defensible position is that **AI training should not be presumed either automatically infringing or automatically exempt**. Its legality should depend upon the nature and source of the copyrighted material, the purpose and scale of use, authorization or licensing, the applicability of Section 52 and the effect of the activity on the legitimate interests of copyright owners.

COPYRIGHT INFRINGEMENT THROUGH AI TRAINING

The use of copyrighted material for training generative AI raises a fundamental conflict between technological innovation and the exclusive rights granted to creators. AI developers may collect enormous quantities of online content, reproduce those works within datasets, process them during model training and subsequently deploy the trained models commercially. WIPO has identified copyright concerns at each of these stages scraping, training and outputs and notes that the legal position remains uncertain across jurisdictions. In India, the principal provisions are Sections **14, 51 and 52 of the Copyright Act, 1957**. Section 14 defines exclusive copyright rights, Section 51 identifies circumstances constituting infringement, while Section 52 establishes statutory exceptions. Therefore, AI training should neither automatically be characterised as infringement nor automatically treated as exempt; each stage requires separate legal analysis[3].

4.1 Web Scraping and Unauthorized Collection of Copyrighted Works

Web scraping involves automated extraction of information from websites or online repositories. In AI development, scraping can be used to assemble large datasets containing books, articles, photographs, artwork, music, software and other creative works. The fact that material is publicly accessible does not, by itself, eliminate copyright protection.

The legal concern arises when protected works are downloaded, copied or stored without authorization. Section 14 of the Copyright Act grants copyright owners reproduction rights, while Section 51 may apply where a person undertakes an act reserved for the copyright owner without a licence or statutory justification. Consequently, developers should distinguish between **lawfully accessed information** and **lawfully reproduced copyrighted expression**.

Websites may also impose contractual terms restricting automated extraction. Therefore, AI

[3] Sections 14, 51 and 52 of the Copyright Act, 1957 https://share.google/xYLjTS64ONmgGztRe

developers face potentially overlapping copyright, contractual and technological restrictions when constructing datasets. The absence of a specific Indian statutory exception for commercial AI scraping creates considerable uncertainty.

4.2 Reproduction of Copyrighted Works during Training

AI training commonly involves copying works into datasets, transferring them between storage systems, preprocessing them and making further computational copies. These activities are legally significant because Section 14 recognizes reproduction as an exclusive copyright right and, for relevant works, expressly encompasses electronic storage. AI developers may argue that the system does not retain conventional copies of the original works because training converts data into numerical representations or model parameters. Copyright owners, however, may contend that the legally relevant copying occurs **before and during training**, when protected works are actually reproduced and processed. The legal inquiry should therefore focus on the nature and purpose of the copies, their duration, the technology involved and whether a statutory exception or licence applies. The fact that the final model does not contain an easily retrievable copy of every training work does not necessarily resolve whether earlier acts of reproduction were lawful.

4.3 Copyright Infringement through Training Datasets

A training dataset can contain thousands or millions of individual copyrighted works. The presence of such works does not automatically establish infringement; copyright protection attaches to particular protected expression and liability requires identification of an act falling within the owner's exclusive rights without applicable authorization or exception. Section 51 of the Copyright Act provides the principal statutory basis for determining infringement. Where protected works have been reproduced without a licence and the use does not fall within Section 52, potential infringement may arise. A further difficulty is **scale**. Traditional copyright disputes generally concern identifiable acts of copying, whereas AI training may involve enormous datasets assembled through automated processes. This makes it difficult for individual creators to determine whether their works were included, obtain evidence of copying and pursue effective remedies. Dataset provenance and documentation are therefore increasingly important. WIPO identifies dataset vetting, licensed or public-domain training data and other practical measures as ways of reducing infringement risks.

4.4 Memorisation and Substantial Similarity in AI Outputs

An AI model may sometimes reproduce portions of its training material rather than merely learning general patterns. This phenomenon is commonly described as **memorisation**. Where a model generates text, images, code or other expression substantially similar to a protected work, a separate copyright question may arise at the output stage. The critical issue is not whether the AI system was trained on copyrighted material alone, but whether the output reproduces protected expression belonging to a particular copyright owner. A mere similarity of ideas, themes or general style would not necessarily establish infringement because copyright ordinarily protects expression rather than abstract ideas. This distinction is particularly important because AI-generated outputs may range from genuinely new expression to near-verbatim reproduction. The legal assessment should therefore consider the originality and protectability of the source work, the extent of similarity, the nature of the reproduced elements and the circumstances in which the output was generated.

4.5 AI-Generated Outputs and Derivative Works

The generation of an output based upon training data creates another layer of copyright complexity. An output may potentially reproduce protected expression or incorporate substantial elements of an existing work. Where an output amounts to an unauthorized adaptation of a protected work, the copyright owner's adaptation rights may become relevant under Section 14.

However, **not every AI-generated work is automatically a derivative work**. A system may learn general artistic or linguistic patterns without reproducing copyright-protected expression from any particular source. The legal distinction between learning from a work and appropriating its protected expression is therefore essential.

The U.S. Copyright Office has similarly emphasized that existing copyright principles can be applied to AI outputs and that copyrightability depends upon sufficient human authorship rather than the mere use of AI. This demonstrates that the questions of **infringement of existing works** and **copyrightability of new AI outputs** should be analysed separately.

4.6 Liability of AI Developers and Technology Companies

Determining liability requires identifying the entity responsible for collecting training material,

constructing datasets, developing the model and deploying the AI service. Where a company itself authorizes or undertakes unauthorized reproduction, Section 51 may provide a basis for infringement proceedings.

The situation becomes more complicated where developers rely on third-party datasets, open-source material or external data providers. Questions may then arise concerning contractual allocation of responsibility, knowledge of unlawful content, reasonable verification measures and the extent of the company's control over the training process.

AI companies should therefore maintain records concerning the origin of training data, licences obtained, filtering procedures and safeguards against memorisation. WIPO specifically recommends dataset vetting and the use of licensed, public-domain or otherwise compliant training material as risk-reduction measures.

4.7 Defences Based on Fair Use and Fair Dealing

One of the principal legal arguments advanced internationally is that AI training may constitute a permissible use of copyrighted works. However, the applicable doctrine differs between jurisdictions. The United States applies **fair use**, whereas India follows the more specifically enumerated **fair-dealing exceptions under Section 52**.

In India, Section 52(1)(a) permits fair dealing with qualifying works for specified purposes, including private or personal use such as research, criticism or review. The difficulty is that large-scale commercial AI training does not expressly appear among these purposes. A commercial developer therefore cannot safely assume that every training activity is protected merely because the activity has technological or research-related characteristics.

Internationally, the debate remains unsettled. The U.S. Copyright Office has examined whether copyrighted works may lawfully be used for AI training and has identified disagreement concerning dataset collection, training practices and permission requirements. WIPO likewise observes that copyright owners and AI developers hold competing positions regarding whether training constitutes infringement or falls within existing exceptions.

Accordingly, the strongest legal approach is to assess **purpose, source of the material, authorization, scale of copying, nature of the protected work, statutory exceptions and the effect on legitimate creator interests** rather than adopting an automatic rule either

permitting or prohibiting AI training. This approach preserves room for legitimate technological development while recognizing that large-scale commercial exploitation of protected creative works may require stronger legal safeguards.

JUDICIAL AND COMPARATIVE PERSPECTIVES

The legal treatment of copyrighted works used for AI training is developing through a combination of **judicial decisions, legislative measures and regulatory initiatives** rather than through a single universally accepted rule. The United States has largely addressed the issue through copyright litigation and the doctrine of fair use, while the European Union has adopted a more structured legislative approach through its copyright framework and the EU Artificial Intelligence Act. These approaches demonstrate that the central dispute is not simply whether AI training involves copying, but whether such copying is legally justified, whether it affects existing and potential markets for creators and what obligations should be imposed on AI developers. The comparison is particularly relevant to India because the Copyright Act, 1957 presently contains no express, comprehensive regime specifically governing commercial AI training.

5.1 Emerging Judicial Developments in the United States

The United States has become a major jurisdiction for litigation concerning the use of copyrighted material in AI training. The principal legal framework is **17 U.S.C. §107**, which establishes the four-factor fair-use test: purpose and character of the use, nature of the copyrighted work, amount used and effect upon the potential market.

A significant development occurred in **Thomson Reuters Enterprise Centre GmbH v Ross Intelligence Inc.,**[4]. The dispute concerned Ross Intelligence's use of Westlaw-related material in developing a competing legal-research AI system. The court rejected Ross's fair-use defence and found that the commercial and competitive character of the use, together with the effect upon the relevant market, weighed against fair use.

The decision is important because it demonstrates that courts may examine **commercial competition and market substitution** rather than treating AI training as automatically transformative. However, it should not be interpreted as establishing that all AI training is

[4] Thomson Reuters Enterprise Centre GmbH v Ross Intelligence Inc., 2025 WL 458520 (D. Del. Feb. 11, 2025)

unlawful. The factual circumstances of each dataset, purpose and market remain important.

The U.S. Copyright Office has also undertaken an extensive examination of AI and copyright. Its Part 3 report specifically addresses generative-AI training and considers the legal implications of using copyrighted works to develop AI models.

5.2 Copyright Litigation Involving Generative AI

A growing body of litigation has been brought by authors, publishers, visual artists, news organisations and other copyright owners against AI developers. Cases involving **OpenAI, Microsoft, Meta and Anthropic** raise questions concerning whether copyrighted works were copied without authorization and whether the resulting AI systems compete with the markets from which the training material originated. Multiple OpenAI-related copyright actions, including claims brought by prominent authors and *The New York Times*, have been consolidated for federal proceedings in New York.

In **Kadrey v Meta Platforms, Inc., No.**[5], the court addressed authors' claims concerning Meta's use of books for AI training. The court granted Meta summary judgment on the fair-use issue based on the evidentiary record presented, while recognising the significance of potential market harm. The decision illustrates the importance of evidence concerning whether AI systems may diminish the economic value of copyrighted works.

The contrasting outcomes in AI-related cases demonstrate that **there is no universal judicial rule that training on copyrighted material is either always fair use or always infringement**. Courts continue to examine the particular facts, including the source of the works, purpose of copying, commercial use, availability of licensing markets and evidence of market harm.

5.3 European Union Approach

The European Union has adopted a comparatively structured legislative approach. A major development is **Directive (EU) 2019/790 on Copyright and Related Rights in the Digital Single Market**, particularly Articles 3 and 4 concerning text and data mining (TDM).[6]

[5] Kadrey v Meta Platforms, Inc., No. 3:23-cv-03417 (N.D. Cal. June 25, 2025)
[6] Directive (EU) 2019/790 on Copyright and Related Rights in the Digital Single Market, particularly Articles 3

Article 3 provides an exception for reproductions and extractions by research organisations and cultural heritage institutions for scientific research purposes, subject to specified conditions. Article 4 establishes a broader TDM exception but permits rights holders to reserve their rights in an appropriate manner, particularly where works are made publicly available online.

This creates a distinctive European model: AI developers may benefit from statutory permission for certain forms of data mining, but the legal framework also recognises the interests of rights holders. The EU approach therefore attempts to create a balance between **data-driven innovation and copyright control**. The model is particularly relevant to India because it demonstrates how legislation can address AI-related data processing without simply eliminating copyright protection.

5.4 Text and Data Mining Exceptions

Text and Data Mining refers broadly to automated computational analysis of large quantities of digital material to identify patterns, trends and relationships. It is highly relevant to AI because machine-learning systems depend upon large-scale computational analysis of data.

Under **Article 3 of Directive (EU) 2019/790**, certain reproductions and extractions made for scientific research by research organisations and cultural heritage institutions are permitted, provided the statutory conditions are satisfied. Article 4 extends TDM permissions more broadly but allows rights holders to expressly reserve their rights.

This **rights-reservation mechanism** is particularly significant. Rather than requiring individual permission for every work before computational analysis, the EU framework permits certain data mining while allowing creators and rights holders to communicate that their works are not available for such use.

The model therefore seeks to reduce transaction costs while preserving an element of creator autonomy. It also demonstrates that a modern copyright framework can differentiate between **non-commercial research, commercial data mining and rights-holder control** rather than applying one rule to every AI-training activity.

and 4 https://share.google/YBaTIhlyPN6udP4dW

5.5 Copyright Obligations under the EU AI Act

The **EU Artificial Intelligence Act, Regulation (EU) 2024/1689**, adds another layer of obligations for providers of general-purpose AI models. Article 53 requires such providers to establish a policy to comply with EU copyright law and related rights, including respecting rights reservations made under Article 4(3) of Directive (EU) 2019/790. Providers must also make publicly available a sufficiently detailed summary of the content used for training their general-purpose AI models. This approach is significant because it introduces **transparency as a regulatory obligation**. Copyright owners cannot effectively exercise their rights if they have no meaningful information concerning what material has been used to train AI systems[7].

The EU model consequently moves beyond traditional infringement litigation and introduces a degree of **ex ante compliance**. AI developers are expected to consider copyright compliance during model development rather than waiting until a creator initiates legal proceedings. For India, this approach offers a potentially valuable policy lesson: transparency concerning training datasets could make licensing, rights enforcement and compensation more practical.

5.6 International Approaches to Creator Protection

International approaches reveal three broad regulatory strategies. The first is the **litigation-led model**, exemplified by the United States, where courts apply established copyright doctrines such as fair use to emerging AI disputes. The second is the **statutory exception model**, represented strongly by the EU's text-and-data-mining framework, where legislation expressly permits specified forms of computational analysis while preserving certain rights of copyright owners. The third is the **transparency and accountability model**, increasingly reflected in AI regulation, where developers are required to document or disclose aspects of their training practices.

The United Kingdom is currently considering alternatives ranging from stronger licensing requirements to a broad data-mining exception and a data-mining exception combined with a rights-reservation mechanism. Its March 2026 impact assessment specifically evaluates these three policy options, demonstrating that the balance between AI innovation and creator

[7] EU Artificial Intelligence Act, Regulation (EU) 2024/1689 Article 4(3) and Article 53 https://share.google/RTDOurzUDStETJaSS

protection remains under active consideration.

For India, these international experiences suggest that a purely litigation-based solution may be insufficient. A sustainable framework could combine **clear rules on AI training, transparency regarding training datasets, licensing opportunities, effective rights-reservation mechanisms and appropriate remuneration for creators**. Such a framework would provide greater certainty to both copyright owners and AI developers while recognising that technological progress should not depend upon the unrestricted commercial use of human creative works.

LEGAL CHALLENGES AND PROTECTION OF CREATORS

The rapid expansion of generative AI has created a significant gap between the **existing copyright framework and the realities of large-scale AI training**. Copyright law is designed around identifiable acts of reproduction, licensing and exploitation, whereas AI development may involve the automated processing of millions of works across multiple jurisdictions. WIPO has recognised continuing uncertainty concerning the use of protected works for AI training and has highlighted transparency, licensing, attribution and compensation as central issues for creators. In India, the challenge is particularly important because the Copyright Act, 1957 contains no express provision specifically regulating commercial generative-AI training. The objective should therefore be to create a framework that protects creators without unnecessarily restricting legitimate technological development.

6.1 Absence of Specific AI-Training Provisions in Indian Copyright Law

The **Copyright Act, 1957** remains the principal legislation governing copyright protection in India, but it was enacted before the emergence of generative AI and does not expressly address AI training datasets, machine learning or text-and-data mining. Section 14 grants copyright owners specified exclusive rights, while Section 51 identifies circumstances constituting infringement. Section 52 provides statutory exceptions, including fair dealing for specified purposes.

The difficulty is that none of these provisions directly answers whether a commercial AI developer may reproduce millions of copyrighted works for training purposes. Consequently, applying conventional copyright principles to AI training may produce uncertainty for both

creators and developers. A specific legislative framework could clarify when permission is required, what forms of computational use are permitted and what obligations apply to AI companies.

6.2 Uncertainty Regarding Fair Dealing

Section 52(1)(a) permits fair dealing with qualifying works for specified purposes, including private or personal use such as research, criticism or review. However, the provision does not expressly mention **artificial intelligence, machine learning or commercial AI training**.

This creates an important distinction between a university using a limited dataset for genuine research and a commercial corporation using vast quantities of copyrighted works to develop a profitable AI model. The commercial purpose, scale of copying, nature of the works and potential impact on copyright markets may all become relevant.

Accordingly, it would be inappropriate to assume that every AI-training activity constitutes fair dealing merely because the technology involves research or automated analysis. At the same time, a blanket rule treating every computational use as infringement could unnecessarily restrict scientific and technological innovation. Clear statutory guidance is therefore required.

6.3 Lack of Transparency in Training Datasets

Creators often face a fundamental evidentiary problem: **they may not know whether their works have been used to train an AI model**. Training datasets can contain enormous quantities of material collected from multiple sources, making individual verification extremely difficult. WIPO has identified transparency as an important component of effective copyright infrastructure and has emphasised the growing need for systems that enable rights holders to understand how their works are being used.

A regulatory framework could therefore require AI developers to maintain appropriate records concerning the provenance of training material and provide meaningful information about dataset categories and sources. Transparency need not require disclosure of confidential technical information or commercially sensitive model architecture. Instead, it should provide sufficient information for creators to identify potential use and pursue licensing or legal remedies where appropriate.

Without meaningful transparency, copyright enforcement becomes disproportionately difficult because the creator carries the burden of discovering an otherwise invisible use of their work.

6.4 Attribution and Recognition of Creators

Attribution is an important component of the relationship between creators and their works. AI systems may process creative works without preserving information about the original author, while outputs may subsequently circulate without any indication of the sources that influenced the model.

This raises both legal and ethical concerns. Although attribution rights under Indian copyright law depend upon the nature of the work and the applicable statutory provisions, creators have a legitimate interest in maintaining recognition of their creative contribution. The issue becomes particularly important where AI-generated content commercially benefits from patterns or expression derived from identifiable creative communities.

A future regulatory framework could encourage **machine-readable attribution metadata, creator identifiers and provenance mechanisms**. WIPO's recent work on AI and copyright infrastructure specifically highlights identification, attribution, metadata and rights-management systems as important tools for protecting creators.

6.5 Compensation and Licensing Mechanisms

The most significant economic concern is whether creators should receive remuneration when their copyrighted works are commercially used to develop AI systems. A licensing framework could allow AI developers to obtain permission from individual copyright owners, publishers, collective management organisations or other rights-management bodies.

Individual licensing, however, may become impractical where training involves millions of works. Therefore, **collective licensing, extended licensing arrangements or standardised AI-training licences** could provide more workable solutions. Such systems could establish transparent royalty mechanisms based on the nature and scale of use.

WIPO discussions have identified **consent, remuneration and transparency** as important principles in the debate concerning creator protection. Licensing would also benefit AI developers by providing greater legal certainty and reducing the risk of extensive copyright

disputes.

6.6 Balancing Innovation with Copyright Protection

The objective of reform should not be to prevent AI development or to permit unrestricted exploitation of creative works. A sustainable copyright framework must recognise that **innovation and creator protection are mutually dependent rather than inherently contradictory**. AI developers require access to high-quality data, while creators require meaningful control and economic recognition when their works are commercially exploited.

A balanced Indian model could therefore combine **transparent training-data practices, licensing mechanisms, reasonable rights-reservation or opt-out systems, collective remuneration and safeguards for legitimate research**. Developers should also be encouraged to use licensed, public-domain or otherwise compliant datasets and implement technical measures to reduce the risk of reproducing protected works. WIPO recommends dataset vetting and verification of ownership and licensing as practical risk-management measures.

The protection of creators should move beyond a purely reactive infringement model. A modern framework should enable creators to **know when their works are being used, exercise meaningful choices regarding that use, receive appropriate compensation where commercially justified and enforce their rights effectively**, while allowing responsible AI innovation to continue. This balance would provide greater certainty to both copyright owners and technology developers and would better align Indian copyright law with the realities of the digital creative economy.

CONCLUSION

The increasing use of copyrighted works for training generative artificial intelligence has created a significant challenge for contemporary copyright law. This study demonstrates that the existing Indian framework under the **Copyright Act, 1957**, particularly Sections 14, 51 and 52, provides a foundation for protecting creators but does not specifically address the complexities of large-scale AI training. The absence of clear provisions concerning training datasets, transparency, licensing and remuneration creates uncertainty for both copyright owners and AI developers. Comparative developments in the United States and European

Union further demonstrate that neither unrestricted access to copyrighted works nor an absolute prohibition on AI training provides a sustainable solution. India therefore requires a balanced framework that recognises technological advancement while preserving the legitimate economic and moral interests of creators. Greater transparency regarding training datasets, appropriate licensing mechanisms, effective rights-reservation options, collective remuneration systems and safeguards against unauthorised reproduction can strengthen creator protection. At the same time, legitimate research and responsible technological innovation should remain protected. Ultimately, copyright law must evolve from a purely reactive model of infringement towards a framework based on **transparency, accountability, consent and fair compensation**. Such an approach would ensure that the growth of artificial intelligence contributes to digital innovation without undermining the creative labour and intellectual contributions upon which the AI ecosystem itself increasingly depends.

REFERENCES

1. India. The Copyright Act, 1957. Act No. 14 of 1957. New Delhi: Government of India; 1957.
2. India. The Copyright Rules, 2013. New Delhi: Ministry of Commerce and Industry, Department for Promotion of Industry and Internal Trade; 2013.
3. World Intellectual Property Organization. Generative AI: Navigating Intellectual Property. Geneva: WIPO; 2024.
4. World Intellectual Property Organization. Generative AI and Intellectual Property: WIPO Conversation on Intellectual Property and Artificial Intelligence. Geneva: WIPO; 2024.
5. United States Copyright Office. Copyright and Artificial Intelligence, Part 3: Generative AI Training. Washington, DC: United States Copyright Office; 2025.
6. Brauneis R. Copyright and the training of human authors and generative machines. Columbia J Law Arts. 2025;48(1):1-59.
7. Buick A. Copyright and AI training data—transparency to the rescue? J Intellect Prop Law Pract. 2025;20(3):182-192.
8. de la Durantaye K. Control and compensation: a comparative analysis of copyright exceptions for training generative AI. IIC Int Rev Intellect Prop Compet Law. 2025; 56:737-770.
9. Dornis TW, Stober S. Generative AI training and copyright law. J Intellect Prop Law Pract. 2025;20(6):
10. Lee E. The Copyright Implications of Generative AI. Berkeley: University of California, Berkeley School of Law; 2024.
11. Desai DR, Riedl M. Between copyright and computer science: the law and ethics of generative AI. Hastings Law J. 2024;75:
12. Yang SA, Zhang AH. Generative AI and copyright: a dynamic perspective. 2024.
13. Ren J, Xu H, He P, Cui Y, Zeng S, Zhang J, et al. Copyright protection in generative AI: a technical perspective. 2024.
14. WIPO. Copyright and Artificial Intelligence. Geneva: World Intellectual Property Organization; 2025.
15. European Union. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market. Official Journal of the European Union. 2019;L130:92-125.

16. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. 2024;L1689.

17. European Union. Regulation (EU) 2024/1689, Article 53: Obligations for Providers of General-Purpose AI Models. Brussels: European Union; 2024.

18. United Kingdom Department for Science, Innovation and Technology; Department for Culture, Media and Sport; Intellectual Property Office. Report on Copyright and Artificial Intelligence. London: The Stationery Office; 2026. ISBN 978-1-5286-6308-3.

19. United Kingdom Department for Science, Innovation and Technology; Department for Culture, Media and Sport; Intellectual Property Office. Copyright and Artificial Intelligence: Impact Assessment. London: The Stationery Office; 2026. ISBN 978-1-5286-6317-5.